

Electronic Theses and Dissertations

2020

Predicting breast cancer progression by using Cell-free DNA.

Bwire, Albert
Faculty of Information Technology
Strathmore University

Recommended Citation

Bwire, A. (2020). *Predicting breast cancer progression by using cell-free DNA* [Thesis, Strathmore University].
<http://hdl.handle.net/11071/12028>

Follow this and additional works at: <http://hdl.handle.net/11071/12028>

Predicting Breast Cancer Progression by using Cell-free DNA

By



Thesis Submitted to the Faculty of Information in partial fulfilment of the requirements for the award of Master of Science in Information Technology

Master of Science in Information Technology

Strathmore University

November 2020

Declaration

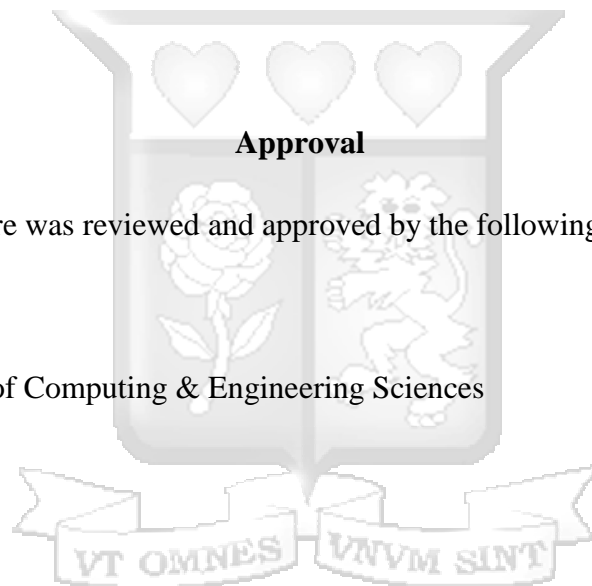
I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

© No part of this thesis may be reproduced without the permission of the author and Strathmore University

Albert Bwire

Signature

Date



The thesis of Albert Bwire was reviewed and approved by the following:

Dr. Joseph Orero, PhD

Senior Lecturer, School of Computing & Engineering Sciences

Strathmore University

Dr. Joseph Orero, PhD

Dean, Faculty of Information Technology

Strathmore University

Dr. Bernard Shibwabo,

Director of Graduate Studies,

Strathmore University

Abstract

Cancer is among the leading causes of deaths in Kenya after infectious and cardiovascular diseases. Among the various forms of cancer, breast cancer accounts for a significant percentage of all new cancer incidences in the country and has a high mortality rate. On a global level, breast cancer is considered the most common cancer. Treatment methods employed vary from patient to patient due to factors such as the stage, age, and health. Treatment methods such as surgery, radiotherapy, chemotherapy or a combination of all have been used all to varying degrees of success and are not always efficient. However, these modalities have been employed successfully when the disease is detected early.

This research applied deep neural networks coupled with genetic algorithms to build a learning model that evaluated the biomarkers obtained from cell-free DNA. The model was able to predict progression of breast cancer. The research, in addition, employed an agile, data-driven methodology due to its recursive nature producing a model with a higher degree of accuracy and specificity. The model developed was able to attain an accuracy of 94% in predicting breast cancer progression.

Keywords: Cell-free DNA, biomarkers, genetic algorithms, specificity

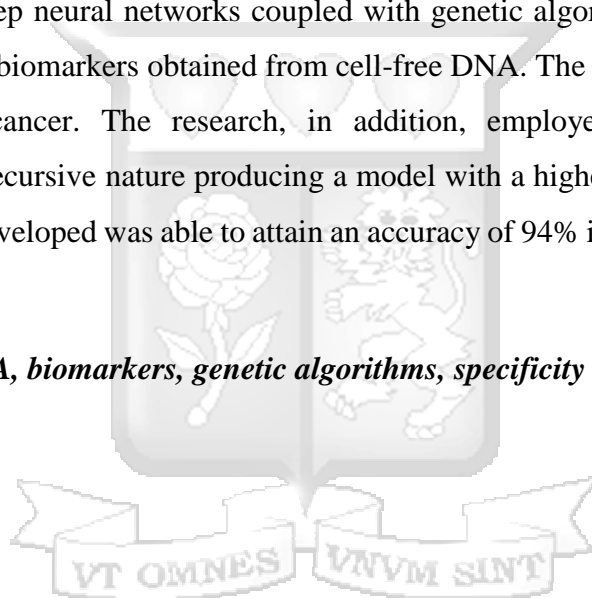
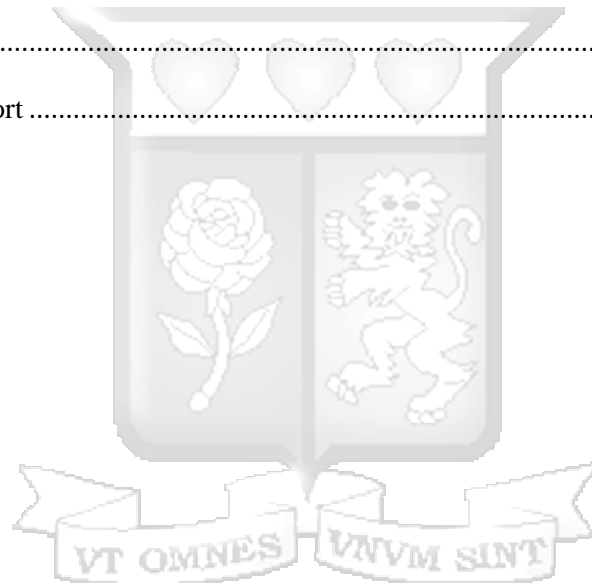


Table of Contents

Declaration.....	ii
Abstract.....	iii
Chapter 1: Introduction.....	1
1.1 Background.....	1
1.2 Problem Statement.....	2
1.3 Aim.....	2
1.4 Specific Objectives.....	3
1.5 Research Questions.....	3
1.6 Justification.....	3
1.7 Scope and Limitation.....	4
Chapter 2: Literature Review.....	5
2.1 Introduction.....	5
2.2 Cancer Monitoring Methods.....	5
2.3 Conventional Cancer Monitoring Methods.....	5
2.4 Non-conventional Cancer Monitoring Methods.....	6
2.4.1 Biopsy.....	6
2.4.2 Liquid Biopsy.....	7
2.5 cfDNA in Cancer Treatment and Monitoring.....	8
2.6 Machine Learning Approach to Predicting Breast Cancer.....	10
2.6.1 The classification prediction problem.....	10
2.6.2 Feature Engineering.....	10
2.6.2.1 Principal Component Analysis.....	10
2.6.2.2 Genetic Algorithms.....	10
2.7 Machine Learning Algorithms.....	11
2.7.1 Recurrent Neural Network.....	13
2.7.2 Random Forest.....	14
2.7.3 Decision Trees.....	14
2.7.4 Naïve Bayes.....	15
2.7.5 Local Interpretable Model-Agnostic Explanations (LIME).....	16
2.8 Related work.....	16

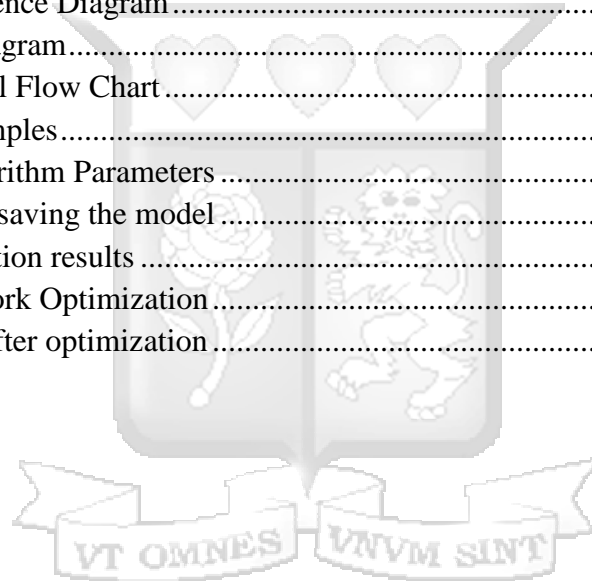
2.9	Conceptual Framework	18
Chapter 3: Research Methodology		19
3.1	Introduction	19
3.2	Research Design	19
3.3	Model Development	19
3.3.1	Data Acquisition	19
3.3.2	Data Processing	20
3.3.3	Model Development	21
3.3.4	Model Validation	21
3.4	System Development Methodology	22
3.5	Research Quality	22
3.6	Ethical Considerations	23
Chapter 4: System Design and Architecture		24
4.1	Introduction	24
4.2	Requirements Analysis	24
4.2.1	Functional Requirements	24
4.2.2	Non-functional Requirements	24
4.2.3	Usability Requirements	24
4.2.4	Reliability Requirements	25
4.3	Architecture	25
4.4	Use Case Diagram	26
4.5	System Sequence Diagram	27
4.6	Sequence Diagram	28
4.7	Flow Chart Diagram	29
Chapter 5: Implementation		31
5.1	Introduction	31
5.2	Extraction of Data	31
5.3	Pre-processing	32
5.4	Training the Model	32
5.5	Validating the Model	34
5.6	Optimizing the Model	35

5.7	Testing the Model	36
5.8	Selecting the Model	37
Chapter 6: Discussion		39
6.1	Introduction.....	39
6.2	Results.....	39
6.3	Research Findings	40
Chapter 7: Conclusions, Recommendations and Future Work		41
7.1	Conclusion	41
7.2	Recommendations.....	41
7.3	Future Work	42
References.....		43
Appendix: Originality Report		49



List of Figures

Figure 2.1: Monitoring breast cancer from cfDNA	9
Figure 2.2: Multi-Layer Perceptron	12
Figure 2.3: Neural network with multiple hidden layers	12
Figure 2.4: Recurrent Neural Network	13
Figure 2.5: Random Forest	16
Figure 2.6: Decision Tree	Error! Bookmark not defined.
Figure 2.7: LIME representation.....	Error! Bookmark not defined.
Figure 2.8: Conceptual Framework	Error! Bookmark not defined.
Figure 3.1: Data-driven modelling.....	22
Figure 4.1: System Architecture	26
Figure 4.2: Breast Cancer Progression Prediction Use Case	27
Figure 4.3: System Sequence Diagram.....	28
Figure 4.4: Sequence Diagram.....	29
Figure 4.5: Trained Model Flow Chart.....	30
Figure 5.1: Extracted samples.....	31
Figure 5.2: Genetic Algorithm Parameters	33
Figure 5.3: Training and saving the model	34
Figure 5.4: Cross-Validation results	35
Figure 5.5: Neural Network Optimization.....	36
Figure 5.6: Test results after optimization.....	37



List of Equations

Equation 2.1: Neuron linear function.....	13
Equation 2.2: Backpropagation.....	13
Equation 2.3: Posterior Probability.....	13



List of Tables

Table 3.1: Data Characteristics 20

Table 5.2: Model results 37

Table 6.1 : Performance Comparison 39



Chapter 1: Introduction

1.1 Background

Cancer is a disease that is characterized by the continuous growth and abnormal uncontrollable division of body cells thereby forming a tumour. The unusual growth can then invade other parts and organs of the host's body eventually leading to death (Cancer, 2018). There are different types of cancer such as prostate, cervical, lung, breast, stomach and many others that afflict different organs of the human body. Cancer is among the leading causes of death in the world accounting for 9.6 million deaths with 70% of the cancer-related deaths occurring in low- and middle-income countries. The prevalence of cancer in low- and middle- income countries has been exacerbated by the risk factors present in these countries such as excessive consumption of alcohol, tobacco use, genetic predispositions, and infections.

Breast cancer is the most commonly occurring cancer in women and the second most commonly occurring cancer overall according to the world cancer research fund international. It is estimated that there are 2 million cases worldwide yearly (Breast Cancer Statistics, 2018). In Kenya, evidence shows that cancer is third leading cause of death after infectious diseases and cardiovascular conditions accounting for 28,000 and an average of 37,000 new cases of cancer (Kenya National Strategy for the Prevention and Control Non-Communicable Diseases 2015-2020, 2015). The situation is getting worse as the number of new cases increase yearly. This necessitates a proactive approach to detection, monitoring and treatment of cancer.

According to Shulman et al., (2010), more than half of the incident cases of breast cancer occur in developing countries. A high proportion of these cases are detected in late stages due to the inadequate facilities, lack of awareness and access to treatment. Additionally, the facilities required are expensive and therefore, in the developing countries, there is a dearth of affordable, high-quality treatment options thereby exacerbating the breast cancer burden in the countries. However, these developing countries realize the urgency of the disease and most countries have strategies to aid in curbing the disease.

There is an emphasis on early detection as cancer can be treated successfully when detected early. However, the threat of recurrence remain an unsolved issue as there are numerous cases of cancer

recurrence. Well developed monitoring mechanisms thereby become vital in the fight against cancer. Constant monitoring can identify recurrence, metastasis and even resistance to treatment. In addition, accountability of artificial intelligence used in healthcare is a major problem (Panch, Szolovits, & Atun, 2018). Deep learning methods have hidden layers through which data is processed to derive an inference. The operations of these layers are abstracted from the users and the results obtained as a result are not describable to a human observer. This research tackled this issue by employing the LIME (Local Interpretable Model-agnostic Explanations) model.

1.2 Problem Statement

Current breast cancer monitoring methods employed are invasive. Monitoring response to treatment is not always efficient due to the nature of the methods employed and their limitations that lead to extended periods of time between testing and clinical application of results (Sobhani, Generali, Zanconati, Bortul, & Scaggiante, 2018). Predictions rely on data retrieved as a result of the monitoring methods used. This research employed an approach to predicting breast cancer progression by combining critical biomarkers of breast cancer derived from liquid biopsies that enumerated circulating cell-free DNA (cfDNA) and other breast cancer biomarkers. The concentration of cfDNA is higher in cancer patients and thus is a useful biomarker in monitoring cancer and as a result in predicting progression.

1.3 Aim

The aim of this research was to develop a model for analyzing the concentration of cfDNA at different points that would be able to predict progression of breast cancer.

1.4 Specific Objectives

- (i) To review the application of machine learning in the prediction of cancer progression
- (ii) To analyze biomarkers used from liquid biopsies to predict disease progression
- (iii) To review the techniques used in monitoring cancer progression
- (iv) To review the use of cfDNA in cancer detection and predicting progression
- (v) To develop and test the model

1.5 Research Questions

- (i) To determine progression of breast cancer, what are the methods employed in monitoring breast cancer?
- (ii) How is breast cancer progression determined?
- (iii) What are the biomarkers used to determine progression of breast cancer?
- (iv) What machine learning techniques have been used to determine progression of breast cancer?

1.6 Justification

According to Cancer (2018), breast cancer accounts for approximately 15% of all cancer deaths among women. Most of these deaths happen in the developing world. The high mortality rate coupled with the heterogeneity of breast cancer tumours necessitates better monitoring. Consequently, the ability to predict, more accurately, how breast cancer will progress will enable further breakthroughs in cancer treatment research. This research therefore builds on previous studies by exploring the co-relation of the different biomarkers in plasma at different cancer stages and derive inferences about cancer progression. Use of predictive analytics used in conjunction with the doctor's skills and input can lead to improved medical care and this research therefore looks to provide a platform to be used by medical professionals in the field of cancer to be able to provide better personalized care by understanding the progression of cancer in a patient's body.

This research took advantage of previous studies that tracked the concentration of cfDNA, drawn from liquid biopsies, as well as genetic expressions of cancer genes, to come up with an adaptive, learning model that can be used to predict breast cancer progression with a higher degree of accuracy.

1.7 Scope and Limitation

This study is limited to working with prior collected data that has been pre-processed. In addition, the study only looked at the co-relation between biomarkers at different stages and seek to derive an inference. Lastly, even though the study can be used to monitor the efficiency of treatment employed, monitoring breast cancer and treatment modalities of breast cancer are outside the scope of this study.



Chapter 2: Literature Review

2.1 Introduction

This chapter covers the importance of monitoring cancer progression and the methods employed. In addition, applications of machine learning that have been employed in predicting cancer, particularly breast cancer, are further analyzed as well as their shortcomings. Lastly, a conceptual framework that will govern this research will be presented.

2.2 Cancer Monitoring Methods

After a diagnosis of cancer, monitoring becomes a crucial process. Depending on the stage, health and wishes of the patient, treatment can commence. The most common cancer treatment methods in the treatment of cancer include surgery such as cryosurgery, laser surgery, radiotherapy, chemotherapy or a combination of the treatment options (Types of Cancer Treatment, 2020). Monitoring, therefore, becomes important as the patients' treatment response can be monitored and potential side effects can be observed and addressed. Furthermore, disease progression can be tracked, and treatment can be administered accordingly. Most importantly, monitoring helps to ensure the patients' health or quality of life.

The key components monitored include the organs affected, the severity of the symptoms, and the side effects of the treatment administered. Ethical considerations in cancer treatment and monitoring need to be observed and patients need to be given a choice and make decisions. Therefore, treatment and monitoring methods employed are dependent on the patient.

There exist different methods employed in monitoring cancer. Conventional methods and non-conventional. These methods are characterized by how widely used they are. Conventional methods have been widely used and include methods such as CT (Computed Tomography) scan, PET (Positron Emission Tomography) scans while non-conventional methods are relatively newer ways of monitoring cancer such as liquid biopsy.

2.3 Conventional Cancer Monitoring Methods

Standard cancer monitoring methods which are widely accepted by the medical research community are in existence and widely used. CT scans use X-rays that generate detailed images and can aid in monitoring the spread of cancer to other organs. During a CT scan X-ray, the machines spin around the patient's entire body, taking pictures at every instant. This gives more

information than static X-rays (Computed Tomography (CT), 2020). The CT scans can help determine whether the tumour is growing bigger or getting smaller. The tumours can be identified even before the symptoms have manifested in the patient's body.

PET scans can be used to emphasize the output obtained from the CT scan. PET/CT scan is a type of nuclear medicine imaging which measures the metabolic activity of cells in body organs. The scan measures the photons that are produced as a result of the annihilation when electrons and positrons in the body combine producing energy. The information obtained is thus used to create images of internal organs. Variations which indicate patient disease state can be detected. This forms the basis of progress monitoring (Nuclear medicine, 2020).

Mammogram, ultrasound and Magnetic Resonance Imaging (MRI) are also used in cancer monitoring. MRI uses an intense magnetic field during the examination process. It produces detailed three-dimensional images. The process takes forty- five to sixty minutes per body part (Magnetic Resonance Imaging (MRI), 2020). During this time, magnets are open on both ends and the patient is expected to lay still. The radiologist can communicate with the patient during the session to instruct about any discomfort and other related concerns.

Despite the success of these conventional imaging approaches, they have limitations. First, they cannot tell if the detected abnormality is a cancer tumour or just a normal scar related tissue. Secondly, they are pivoted around clinical morphology metric where tumour size, shape or colour is measured before, during and after therapy. Thirdly, the response to treatment of patients is not efficiently measured as the size and shape of the tumour are not adequate metrics. Lastly, the output of these techniques is often obtained late during treatment.

2.4 Non-conventional Cancer Monitoring Methods

Non-conventional cancer monitoring methods include biopsy, diffusion-weighted MRI and mobile symptom monitoring.

2.4.1 Biopsy

A biopsy is a technique used to diagnose and monitor cancer. Many argue that it is a sure way of diagnosing cancer. However, this argument fall short as the information obtained from the biopsy is a snap-shot of a tumour and hence does not capture the heterogeneity of tumours and can be difficult to predict early recurrence of cancer (Crowley, Di Nicolantonio, Loupakis, & Bardelli,

2013). The process comprises the removal of physical body tissue for examination followed by imaging using specialised medical devices such as microscopes. The cells are extracted from the sample using a needle and examined for tumours. The procedure varies among patients based on the type and location of the tumour. Apart from diagnosis, it has the potential for monitoring of the treatment response through analysis of genes related to tumours. Crowley et al. (2013), reported that biopsy can be used in determining the disease progression and patient response to the treatment method.

The biopsy process has some challenges associated with it. Firstly, it is invasive. The process is painful for the patient. Secondly, surgical complications exacerbate the chances of cancer spread to other parts of the body. These limitations have significantly inhibited its adoption and as a result, other non-invasive options such as liquid biopsy have gotten much attention by the medical community.

2.4.2 Liquid Biopsy

Liquid biopsy is a non-invasive approach used to detect and monitor cancer. It involves the use of simple blood tests to detect cancer agents such as circulating cell-free DNA (cfDNA) and Circulating Tumour Cells in blood. The latter and former have proven to bear a direct correlation with the presence of tumour cells. Saliva, urine, plasma, cerebrospinal and seminal plasma are other biological fluid variations that can be used (Di Meo, Bartlett, Cheng, Pasic, & Yousef, 2017).

In contrast to biopsy and conventional imaging methods which are dependent on cell tissue and images respectively, liquid biopsy is a tumour marker monitoring method. A blood sample can be used to determine disease and tumour progression. This method overcomes some of the limitations of tissue-based biopsy by allowing for the screening biomarkers which are secreted by tumour cells into the blood. Crowley et al., (2018), reports that liquid biopsy can provide the genetic landscape of all cancerous lesions as well as the opportunity to track genomic evolution. In addition, tracking tumour-associated genes by use of blood can be used to evaluate the presence of residual disease, recurrence, relapse and resistance. The non-invasive nature of liquid biopsy allows for continuous monitoring over extended periods of time.

The general liquid biopsy process is composed of a collection of blood samples, preparation, detection using an imaging device, data analysis and interpretation. DNA based Liquid biopsy

requires 3 ml of blood fluid for plasma preparation. This should be carried out within 5 - 6 hours of withdrawal. The collection is done using tube cleansed with an anticoagulant such as ethylenediaminetetraacetic acid. Cells are thereafter extracted through centrifugation followed by plasma removal after a successful blood clot. The DNA related to tumours are then extracted using the recommended medical kits such as Maxwell RSC (MR), ccfDNA Plasma, Nucleic Acid Isolation, MagNA Pure Compact (MPC), QIAamp Circulating Nucleic Acid (QCNA). Currently, there is no agreed extraction method, but each has associated sensitivity and accuracy issues. Additionally, the method of extraction of cfDNA can also affect the DNA yield and should be considered in the analysis (Pérez-Barrios, et al., 2016).

2.5 cfDNA in Cancer Treatment and Monitoring

Cell-free DNA consists of DNA fragments released after cell death processes from both normal cells and tumour cells (Sobhani et al., 2018). Tumour cells release DNA via multiple mechanisms allowing detection of cancer-associated genetic alterations (Zhang, et al., 2018). Cancer patients, therefore, have a higher cell turnover than normal patients and thus have a higher concentration of cfDNA.

Sobhani et al., (2018) opine that cfDNA obtained from plasma through liquid biopsy tackles the problem of searching for tumour DNA mutations of targeted genes making it effective in monitoring cancer. Obonyo (2018) focused on CTC (circulating tumour cells) obtained from liquid biopsies. However, this approach is limited because CTCs do not always indicate genetically cancerous cells. Thus, cfDNA was used in this study as it is theoretically representative of all tumours and can differentiate between signals from non-cancer and pre-cancerous cells (Sobhani et al., 2018).

Sobhani et al., (2018) determined there is a possibility of monitoring breast cancer by use of cfDNA as shown in figure 1 either qualitatively by looking at mutations or quantitatively. However, there are limitations to both approaches. Qualitative studies done of the mutations only represent approximately 40% of breast cancer patients while quantitative studies of cfDNA in breast cancer patients, cfDNA values can be impacted by viruses or infections which can increase the concentration of cfDNA in the patients.

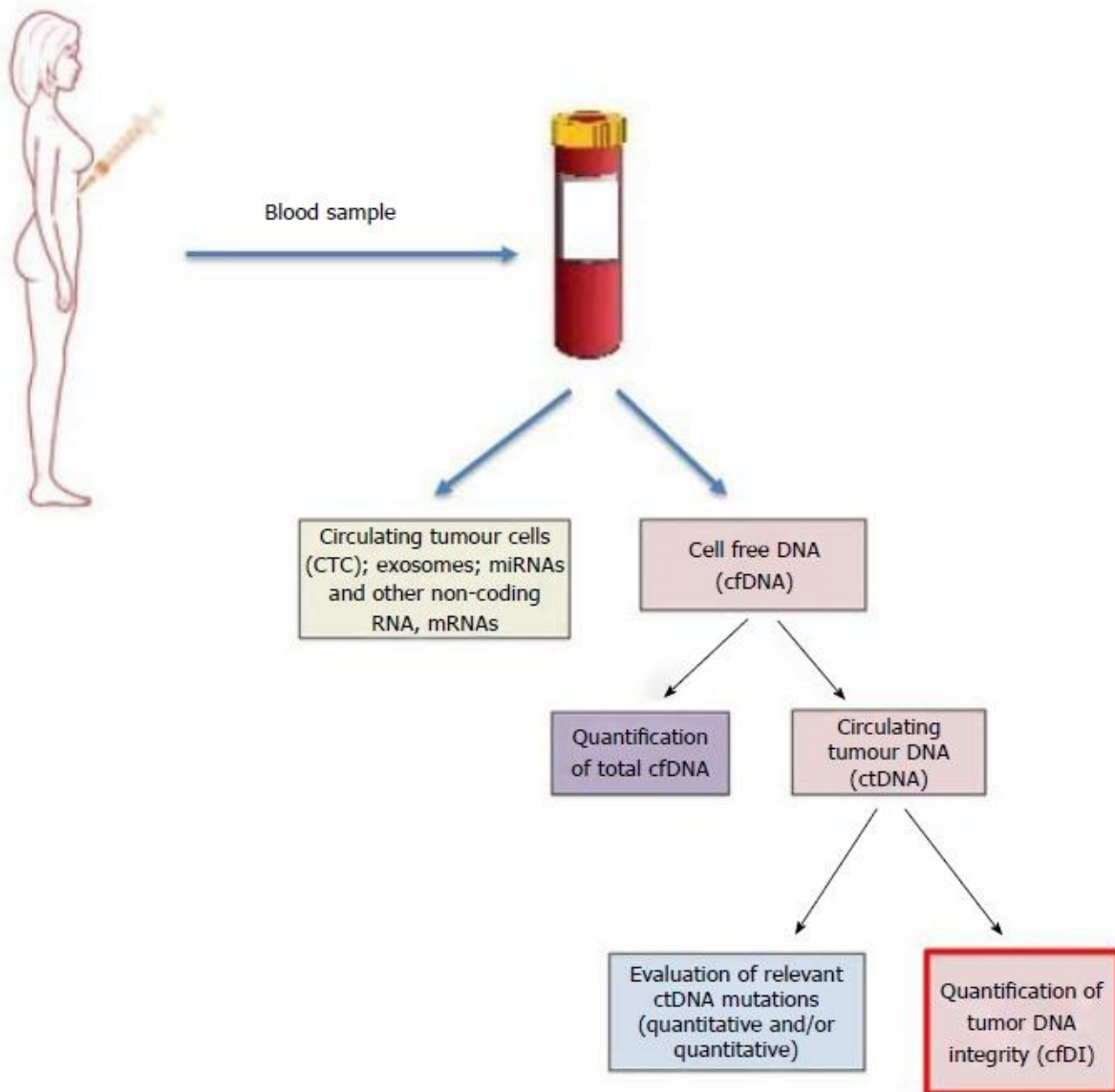


Figure 2.1: Monitoring breast cancer from cfDNA

Nonetheless, cfDNA presents great advantages in the context of breast cancer monitoring. The fact that it can be monitored repeatedly, without having an invasive procedure done on the patient, even after the primary tumour has been removed can aid in the prediction of recurrence (Cheng, et al., 2017). Crowley et al., (2018), reported that monitoring of tumour-associated genetic aberrations through cfDNA obtained from blood can be used to detect the émergence of resistant cancer cells 5-10 months before conventional methods.

2.6 Machine Learning Approach to Predicting Breast Cancer

2.6.1 The classification prediction problem

Classification has been commonly used in previous breast cancer prediction studies to determine whether breast cancer is present or not, whether a tumor is benign or malignant and thus their diagnostic capabilities can prove to be very effective (Amrane, Oukid, Gagaoua, & Ensari, 2018).

Given a training data set, the classification prediction problem is the approximation of a mapping function from the input variables contained in the training data set to output variables. Thereby, a class or category is predicted for a given observation (Brownlee, 2019).

The learning algorithm thus learns a classification function from the training data set that maps the observation(s) to a given category. The classification function is then able to predict previously unseen observations to their respective categories.

2.6.2 Feature Engineering

Feature engineering is a critical step when preparing data with high-dimensionality. It transforms the raw data into features that can be selected that better represent the problem under study. The resultant data is thus suitable for modeling as the computational complexity is reduced and gives rise to an improved model performance (Rençberoğlu, 2019). Principal component analysis and genetic algorithms are two methods that can be used to reduce the dimensionality of data.

2.6.2.1 Principal Component Analysis

Principal component analysis transforms large data sets into smaller data sets by identifying the correlations and patterns while still retaining as much of the variability of the data, thereby preserving most of the data's valuable information. The reduction in the number of variables taken into account for analysis results in faster models due to the reduction in computational complexity (Howley, Madden, O'Connell, & Ryder, 2006). In breast cancer micro array data, principal components can be used to uncover groups of genes that express together and can be useful in breast cancer prediction studies (Bair, Hastie, Debashis, & Tibshirani, 2006).

2.6.2.2 Genetic Algorithms

This is an optimization approach with the goal of simulating evolutionary processes of a living species. It is beneficial when the search space is complex so that the full search is not feasible

through conventional techniques (Niazi & Leardi, 2012). The quality of the possible solutions obtained is evaluated by a fitness solution that is specific to the problem under study. Genetic algorithms thus rely on operators such as crossover, mutation and selection.

2.7 Machine Learning Algorithms

An algorithm is a systematic set of steps taken by a computer program to process a given set of input. This study used genetic algorithms and a deep learning optimized with the stochastic gradient descent algorithm.

Artificial neural networks provide a generalized method for learning different valued functions from examples. The motivation behind the working of these networks has been from observing biological systems (Mitchell, 1997). The general structure of the artificial neural networks is inputs are sent to the hidden layers where the processing is done and output(s) is given. This 3-layer structure is referred to as a multi-layered perceptron as shown in Figure 2.2 (Kim & Kim, 2019).

Neural networks generally consist of connected nodes called neurons whose weights change as the artificial neural network learns. The first component of the neurons is the linear function whose output is the sum of the inputs each multiplied by the weights. The linear function would be expressed as shown in equation 2.1 for input x_1, x_2, x_3 , and output z (Bernico, 2018).

$$z = x_1\theta_1 + x_2\theta_2 + x_3\theta_3 + b$$

Equation 2.1: Neuron linear function

Where $\{\theta_1, \theta_2, \dots, \theta_n\}$ are the weights to be learnt given the input and b is the bias.

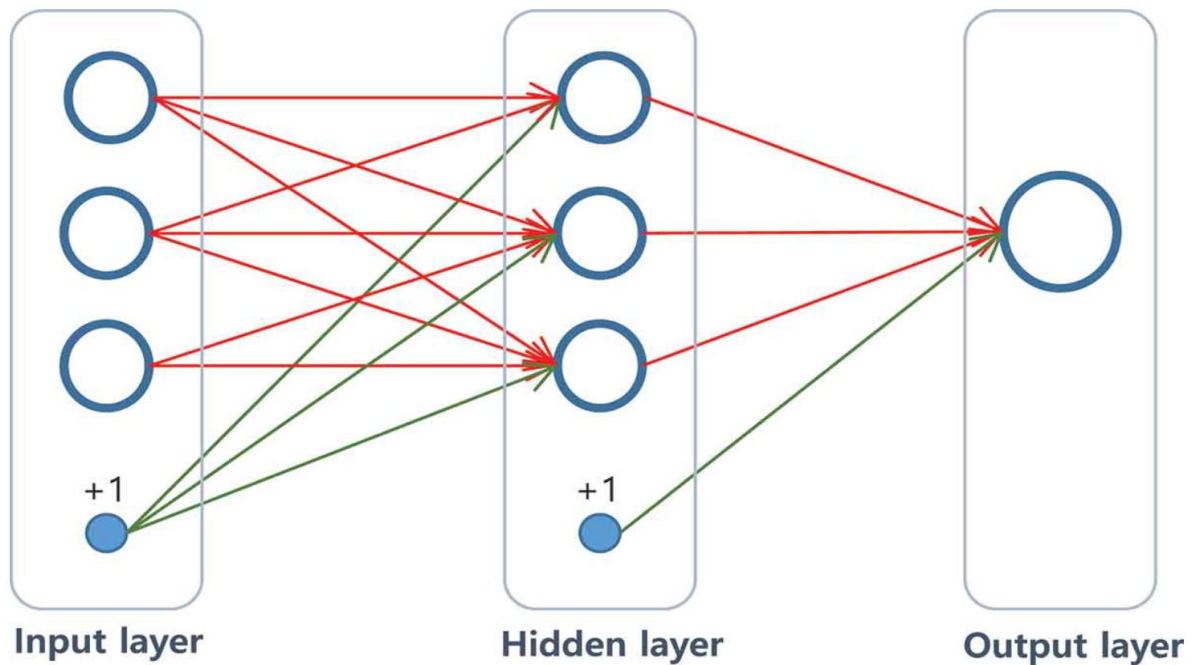


Figure 2.2: Multi-Layer Perceptron

Deep learning which is a form of Artificial Neural Networks is characterized by a neural network made up of multiple hidden layers (Bernico, 2018). Figure 2.4 depicts the structure of a deep learning neural network. Recurrent neural networks is an example of deep learning that was used in this research to come up with the predicted outputs.

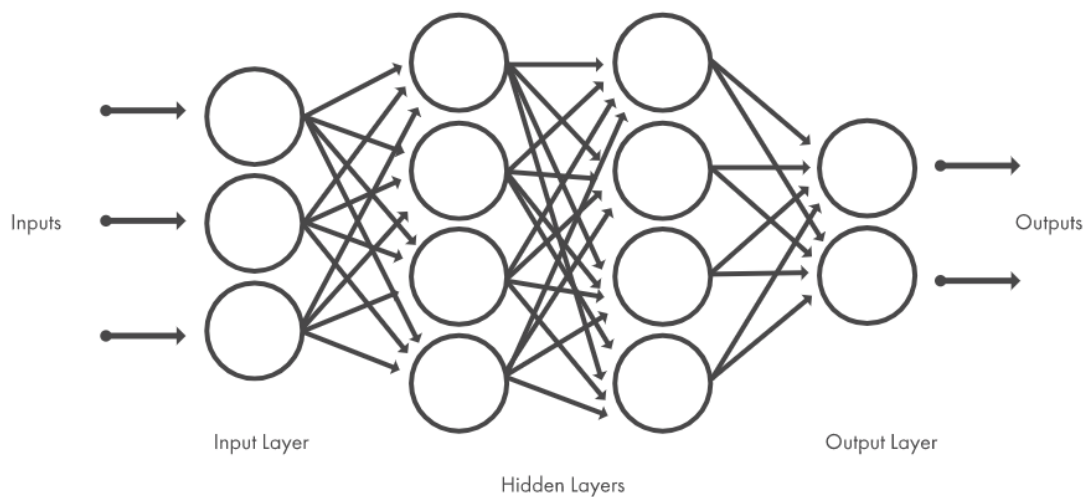


Figure 2.3: Neural network with multiple hidden layers

2.7.1 Recurrent Neural Network

Recurrent neural networks (RNN) are a set of neural networks that allows previous output to be used as input while having hidden layers. It can be represented as shown in Figure 2.2. This algorithm is advantageous as it affords the possibility of processing input of any length. In addition, it takes account of historical information and hence can be useful in processing data where historical information is vital in predicting future characteristics.

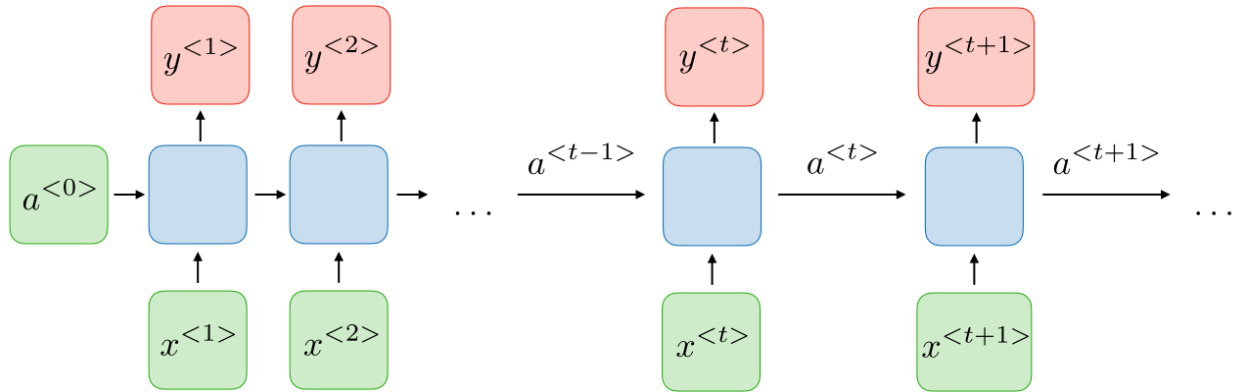


Figure 2.4: Recurrent Neural Network

Recurrent Neural Networks rely on memorization of the information in the sequence. Due to this, it may have potential pitfalls such as being memory-intensive. To reduce the error involved with using the model, backpropagation through time is used and is done at each stage (Amidi & Amidi, n.d). Backpropagation employs gradient descent to attempt to minimize the squared error between the model output values and the targeted values. It can be represented as shown in equation 2.1 where E represents the sum of errors over the output units (Mitchell, 1997).

$$E(\vec{w}) \equiv \frac{1}{2} \sum_{d \in D} \sum_{k \in \text{outputs}} (t_{kd} - o_{kd})^2$$

Equation 2.2: Backpropagation

The gradient descent is the backbone of the backpropagation algorithms as it can serve as the basis of learning algorithms that work through heterogeneous data spaces. The gradient descent is hence the calculation of the steepest descent along the errors (Mitchell, 1997).

2.7.2 Random Forest

Random forests is a classification algorithm that consists of many decision trees thus creating correlated trees. Due to its simplicity and diversity it can be used for both classification and regression tasks. While growing the trees, random forest adds additional randomness by searching for the best feature among a random subset of trees. Lastly, after the random selection of trees and features, the final prediction is determined by majority voting or averages (Ahmad & Yusoff, 2013). Figure 2.3 depicts an example of random forest with two trees.

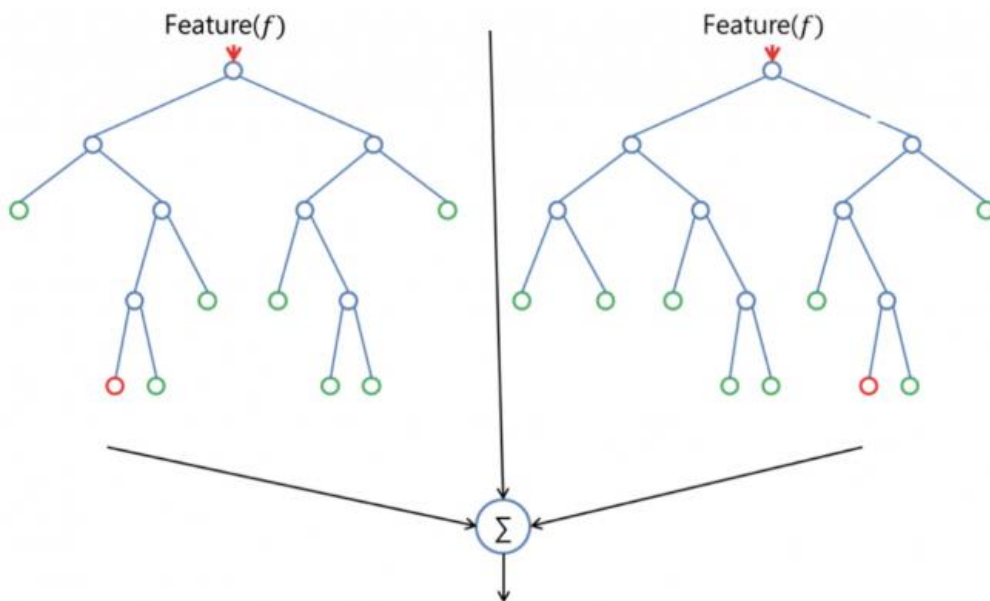


Figure 2.5: Random Forest

2.7.3 Decision Trees

Decision trees are used in classification and regression training and are determined by iteratively partitioning the input from a root node to multiple branch nodes. The root node is the first division from which other nodes are created. The division is based on condition tests such as the gini index, entropy or the classification error that determine the nodes the input are split into. This recursive splitting can continue until the maximum depth of the tree is achieved (Abreu, Santos, Abreu,

Andrade, & Silva, 2016). The result of a decision tree is a set of rules that can be used to make predictions. Figure 2.4 depicts the tree structure.

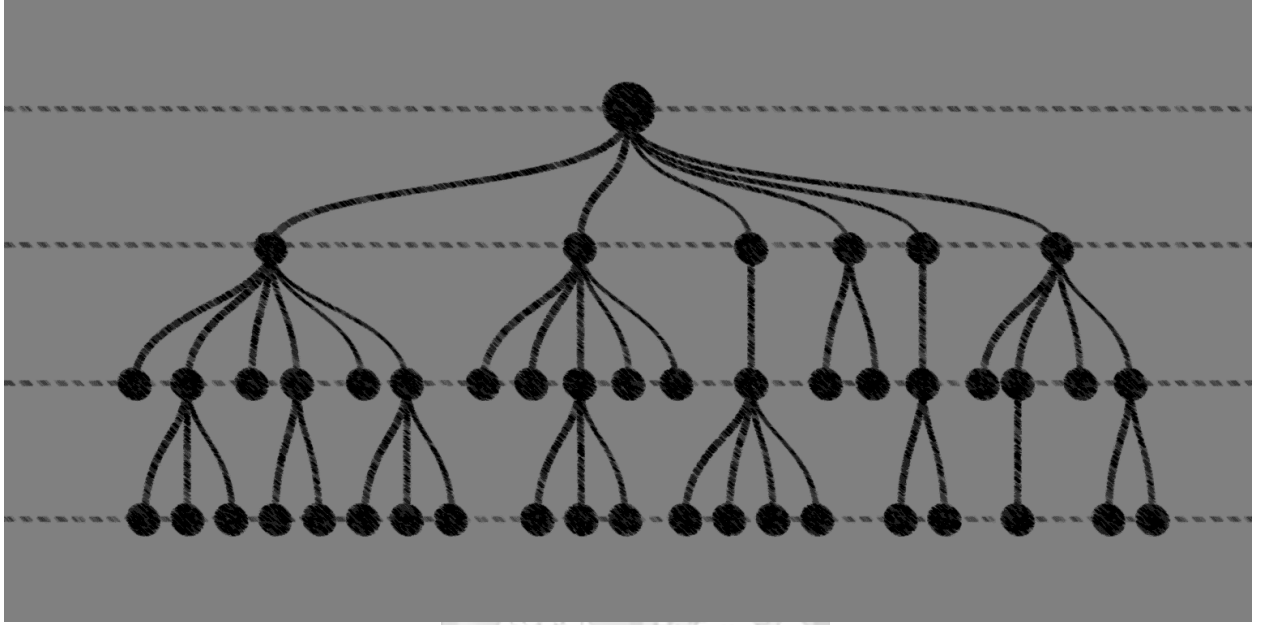


Figure 2.6: Decision Tree

Decision trees have been used in breast cancer studies to predict recurrence and also predict patients' risk of cancer. Decision tree model has been employed in gene selection and classification on high-dimensional microarray data and used to predict risk of breast cancer to a high degree of accuracy (Hamim, El Moudden, Moutachaouik, & Hain, 2020).

2.7.4 Naïve Bayes

In making a decision, the Naive Bayes classifier takes into account the probability distributions of patterns in each class of the data set. The assumption it makes is that there exists a probabilistic relationship between the features and the class. Given a sample S and class C , the learning process involves determining the probability $P(C|S)$. This is the probability that the class belongs to the given sample. The posterior probability $P(S|C)$ is the probability of observing the sample S given the class C (Abreu, Santos, Abreu, Andrade, & Silva, 2016). Using Bayes theorem, the likelihood of a new data sample S_i can be formulated as shown in equation 2.3.

$$P(C|S) = \frac{P(S|C) \times P(C)}{P(S)}$$

Equation 2.3: Posterior probability

2.7.5 Local Interpretable Model-Agnostic Explanations (LIME)

Ribeiro, Singh, and Guestrin (2016) argued that explaining predictions is a vital aspect of getting humans to trust machine learning models and use them effectively. Given that deep learning models operate as a black box, with the operations of the hidden layers not known, effectively adopting them is a challenge especially in exact fields such as medicine. They, therefore, proposed a model that could explain the predictions of any classifier by approximating it with an interpretable model. The model can be illustrated as shown in Figure 2.5. In the representation, the model predicts flu from the inputs and LIME highlights the factors that led to the prediction. The doctor decides whether to trust the prediction.

LIME can be used in breast cancer prediction studies to aid in acceptance of the models generated as well as their validation by oncologists. In addition, the models will benefit from the explanations as errors in the predictions can be highlighted and improved upon.

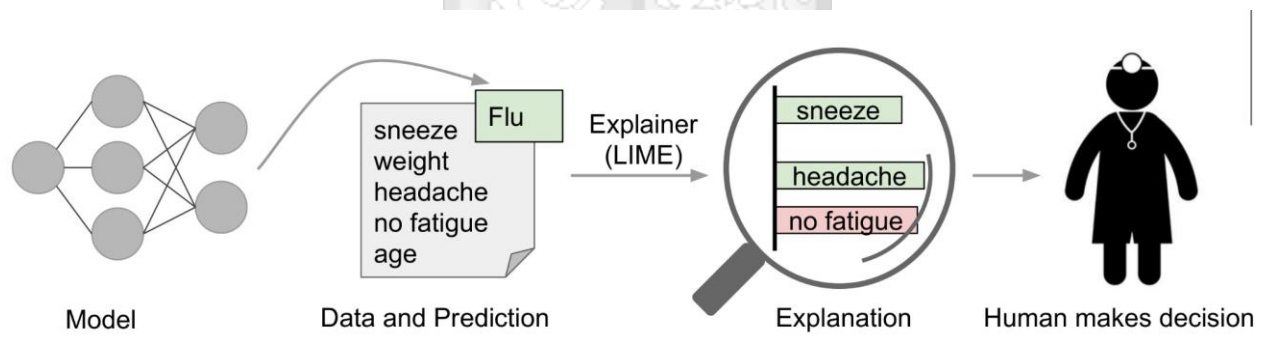


Figure 2.7: LIME Representation Adapted from (Ribeiro, Singh, & Guestrin, 2016)

2.8 Related work

Numerous machine learning algorithms have been employed in numerous studies on breast cancer. Some studies have focused on treatment, monitoring, prediction of recurrence, metastasis, disease progression and the survival rate.

Oketch (2018) used circulating tumors to monitor breast cancer progression. The study focused on monitoring breast cancer progression by using deep learning, specifically convoluted neural

networks to identify the circulating tumor cells. This study was limited to monitoring and thus predictive capabilities were not explored.

Using serum biomarkers determined that the random forest model, with an accuracy of 75%, was the optimal algorithm to predict breast cancer metastasis 3 months in advance. By predicting metastasis early, the study may be helpful in early detection and early treatment (Tseng, et al., 2019). However, the study only evaluated one gene namely; HER2. This presents a limitation as better results can be obtained through the evaluation of a combination of genes.

Some studies automated the process by which recurrences were identified. Using natural language processing, doctors' progress notes were automated (Zeng, et al., 2018). The support vector machine was determined to have the highest accuracy. Nonetheless, this approach does not really predict progression nor provide further insights into breast cancer.

Other studies showed how vital constantly monitoring cancer patients is. A smartphone chatbot application to monitor older cancer patients was used and timely interventions were made to the patients under the study ensuring care was provided promptly (Piau, Crissey, Brechemier, Balardy, & Nourhashemi, 2019).

Moreover, other studies have recognised the importance of having explanations for machine learning based predictions. Case based reasoning approach to explaining predictions is one such example (Lamy, Sekar, Guezennec, Bouaud, & Séroussi, 2019). The cases were retrieved from the database based on the similarity of the query received. However, the drawback of this approach was that the accuracy scores received were approximately 80%.

This study builds upon other studies on employing machine learning algorithms to breast cancer by predicting how breast cancer will progress by comparing the accuracy obtained by different machine learning algorithms.

2.9 Conceptual Framework

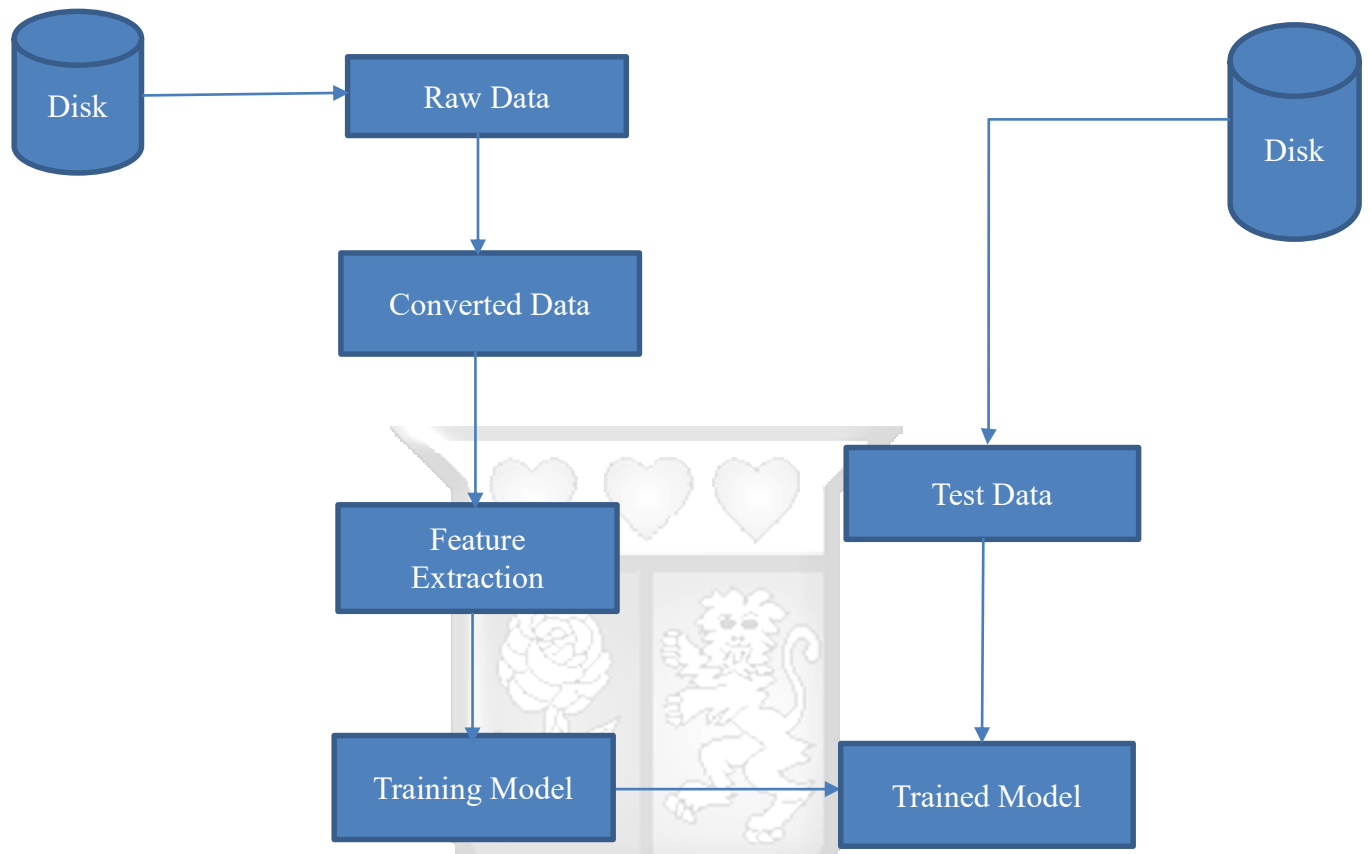


Figure 2.8: Conceptual Framework

Chapter 3: Research Methodology

3.1 Introduction

In this chapter, the research methodology used is outlined. The aspects covered include the approaches to data collection, processing, the system development methodology as well as perspectives on research design, research quality and the ethical considerations adhered to.

3.2 Research Design

Research design is a strategic, systematic, coherent and reasonable way that incorporates all components of the research to completely address the problem under study (Oduor, 2017). Therefore, quality research design implies the application of a well-structured methodology in the collection, measurement and analysis of data to arrive at solutions to the research questions. To arrive at the solutions to the research questions, secondary data analysis was done. This approach was chosen due to the time constraints in completing the research within the set timelines. In addition, due to the specialized data requirements, this approach was the most feasible.

This research is applied research. The aim of this research is to find a solution that can be used by society to alleviate the problems faced. It is for this reason that it is applied research due to the envisaged application of knowledge obtained to solve problems in society.

3.3 Model Development

The model is an artificial neural network that will be developed as follows:

- I. Obtaining data
- II. Processing of data
- III. Extraction of features
- IV. Development of the model
- V. Validation of the model

3.3.1 Data Acquisition

The data was obtained from the publicly available database of the European Biomedical Institute. The data was retrieved from the Array Express database under the number E-MTAB-624. The primary collectors of the data that made the data available publicly adhered to the Declaration of Helsinki and the patients under study gave written and informed consent to participate in the study.

The data is in the form of .CEL files which were downloaded as numerous zip files due to the size of the data. The data consists of information of 65 breast cancer patients and 8 control patients. Of the 65 patients, 15 had a recent diagnosis of breast cancer and 50 had been receiving treatment over a 3 year period (Shaw, et al., 2012). The data which is genotyped by array, thus consisted of the patients' clinical history, their individual genetic characteristics and the disease state. The specific variables in the dataset include the type of surgery undergone by the patients, tumor grades and size, menopausal status and genes. The original work was based on use of CfDNA to infer dormancy as opposed to progression of cancer that this paper focuses on. Table 3.1 shows the characteristics of the data obtained showing the number of patients with different tumour sizes and grade and at different stages of treatment.

Table 3.1: Data Characteristics

	Pre-treatment primary breast cancer	Primary breast cancer following treatment	Metastatic breast cancer
Tumour Size			
T1	43	9	10
T2	24	8	7
T3	8	1	3
T4	0	0	0
Unknown	3	4	10
Tumour Grade			
1	5	3	0
2	31	16	17
3	41	2	9
Unknown	1	1	4

3.3.2 Data Processing

The data is in .CEL file format, therefore, to be able to analyse the data, pre-processing of the data using the dChip which is an open source program was vital. In addition, since the data contains over 200 genomic profiles, factor analysis was done to determine the combination of features that will be studied.

The data was processed by performing normalization and summarization. The output of this process was a .txt file that was used in training and testing of the model. The file contained genes and microarrays that were subjected to further analysis.

On further analysis of the data, data from the 8 control patients were excluded from further processing. The remaining data was retrieved and contained 57 different attributes that included the ER status, disease free months, HER status, tumor stage, size of the tumor, time between diagnoses, time to death, among others for all the patients under study.

3.3.3 Model Development

The model was developed using Weka which is an open source tool. It was chosen due to its capability to develop recurrent neural networks in addition to other neural networks.

The deepLearning4j library by Weka was used to develop a deep learning model from the processed data. The results from the deep learning and other algorithms used to predict progression of breast cancer were used to compare and determine which algorithm is the most accurate for the study.

The model was developed by first, splitting the dataset into three: 60% training, 20% validation and 20% testing. The split into three was intended to make the model better at generalizing with data it has yet to see. Due to the vast number of features in the dataset, feature extraction was performed on the training dataset by employing methods such as the Principal Component Analysis technique and Genetic Algorithms to identify patterns in the data. The intention of deploying these feature extraction techniques was to decrease the features to be studied by either eliminating some or combining some features into one or reducing redundant features in the dataset. The data was then fed into learning algorithms and comparisons done on the accuracy and specificity of the models. The model identified was then validated and tested.

3.3.4 Model Validation

The model was evaluated using the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Relative Absolute Error (RAE), Root Relative Squared Error (RRSE) and the Kappa statistic. In addition, the accuracy of the model against the test data will be a factor in validating the model.

3.4 System Development Methodology

The system development methodology is a structured framework with organized principles and methods that are used to deliver a system (Okonkwo & Huisman, 2018). This research utilizes an agile approach to the data-driven methodology.

The data-driven methodology is based on computational intelligence and machine learning and therefore does not necessitate explicit knowledge of the physical behaviour of the system. It involves the analysis of input, internal and output variables and finding connections between the variables (Solomatine, See, & Abrahart, 2009). Figure 3.1 depicts the methodology.

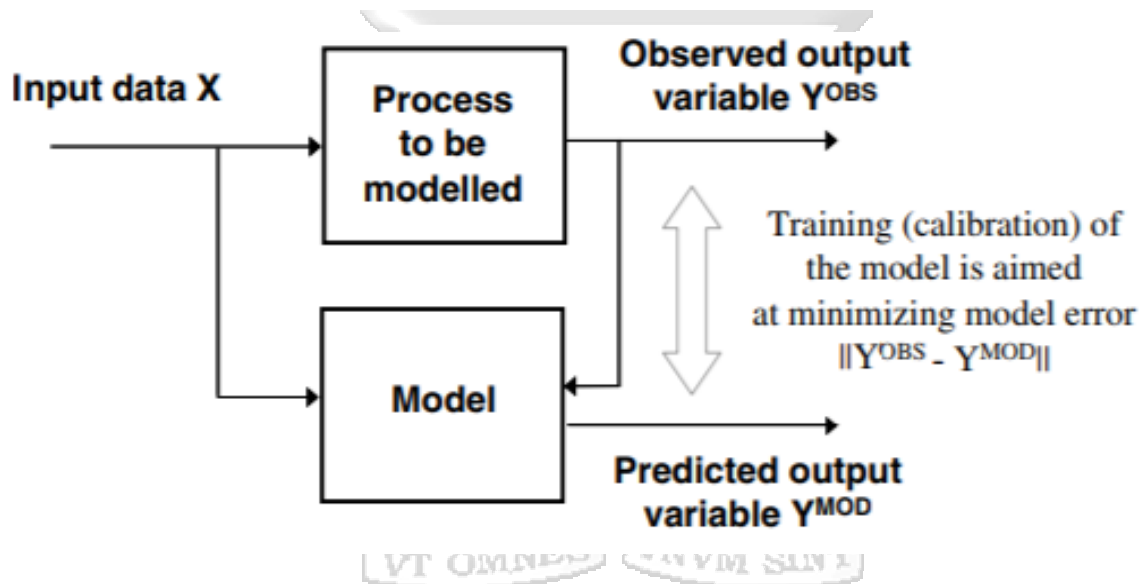


Figure 3.1: Data-driven modelling(Adapted from (Solomatine, See, & Abrahart, 2009))

Agile software development was employed due to its iterative nature. The basic principles of this methodology are working code and working together (Highsmith & Cockburn, 2001). The emphasis on iteration and working code was observed in this research by continuously improving the model to achieve a higher model accuracy. Working together entailed constant interaction with the stakeholders of this research thereby improving the overall quality of the research and the output.

3.5 Research Quality

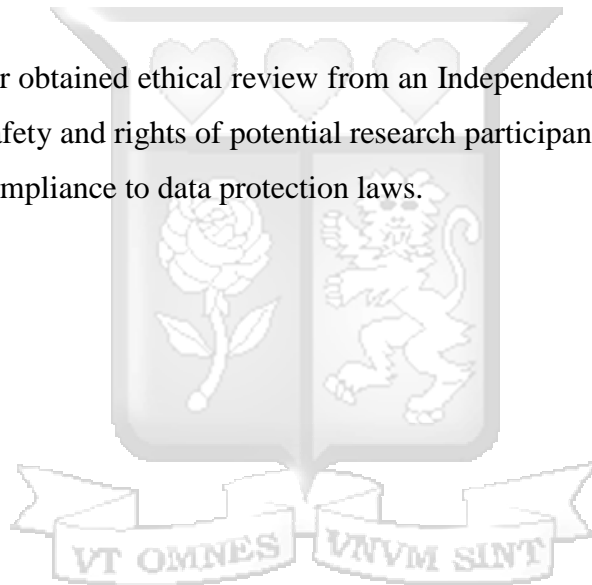
The researcher intends to provide quality research by ensuring that all materials accessed and used in the course of the study are properly cited. In addition, the quality was ensured by obtaining

literature sources from reputable sources. Lastly, to guarantee quality research work, the researcher validated his findings by use of empirical validation methods for neural networks to ensure a higher degree of acceptability of the model.

3.6 Ethical Considerations

This research utilized publicly available data obtained from the European Bioinformatics Institute. The institute has ensured that data that is publicly available only contains de-identified data. Use of the data was credited in this research to the owners of the data. The dataset used contains only data such as gender, age, tumor type, grade, type of procedure, the status of receptor hormones and gene biomarkers which cannot be used to identify a patient and hence ensuring their privacy and confidentiality.

In addition, the researcher obtained ethical review from an Independent Ethical Review Board to aid in ensuring that the safety and rights of potential research participants are safeguarded as well as to help in certifying compliance to data protection laws.



Chapter 4: System Design and Architecture

4.1 Introduction

This section outlines the architecture of the developed prediction model for breast cancer progression. The architecture build up on the conceptual model developed in Figure 2.5. This section covers the requirements that need to be satisfied, the components of the developed system, the interaction between the end user and the developed system. In addition, it covers the interactions between the components of the developed model. Use case diagrams, sequence diagrams, context diagrams as well as data flow diagrams were used to model the system.

4.2 Requirements Analysis

The requirements will be broken down into five namely: Functional requirements, non-functional requirements, usability requirements, reliability requirements and performance requirements.

4.2.1 Functional Requirements

- i. The system should allow the user to input data as CSV, TSV or XLS files. All other forms should be rejected.
- ii. The system should extract the PR, ER and HER2 statuses.
- iii. The system should be able to predict progression of tumor by use of deep learning.
- iv. The predicted progression status should be valid based on the input supplied from the user
- v. The system should provide an explanation on how the progression was arrived at

4.2.2 Non-functional Requirements

- i. The system should be easy to re-train and adjust the weights
- ii. The system should be able to robustly generalize new instances and avoid overfitting
- iii. The system should be secure to avoid unauthorized changes to the model parameters
- iv. The system should have persistent weight storage to avoid re-training every time a prediction is done

4.2.3 Usability Requirements

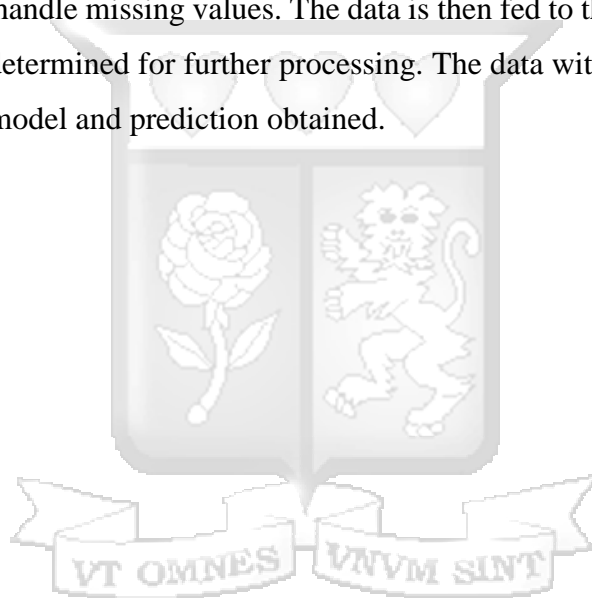
The system is intended to be used in a clinical setting; hospitals and clinics. The system should therefore be simple to use. It should also be straightforward to ensure that it can be easily accepted by the users. The system should also provide accurate predictions and explanations to the predictions as it may directly impact the lives of the patients.

4.2.4 Reliability Requirements

- i. The system should always interface with the existing database containing clinical information
- ii. The administrator should be able to correctly restore the system in the occurrence of a failure

4.3 Architecture

The system architecture gives a generalized layout of the breast cancer progression prediction prototype and its individual components as shown in Figure 4.1. The process begins by extracting data from the database that is supplied by the user. The data is then pre-processed to clean the data, detect outliers as well as handle missing values. The data is then fed to the feature extractor where the suitable features are determined for further processing. The data with the extracted features is then fed into the trained model and prediction obtained.



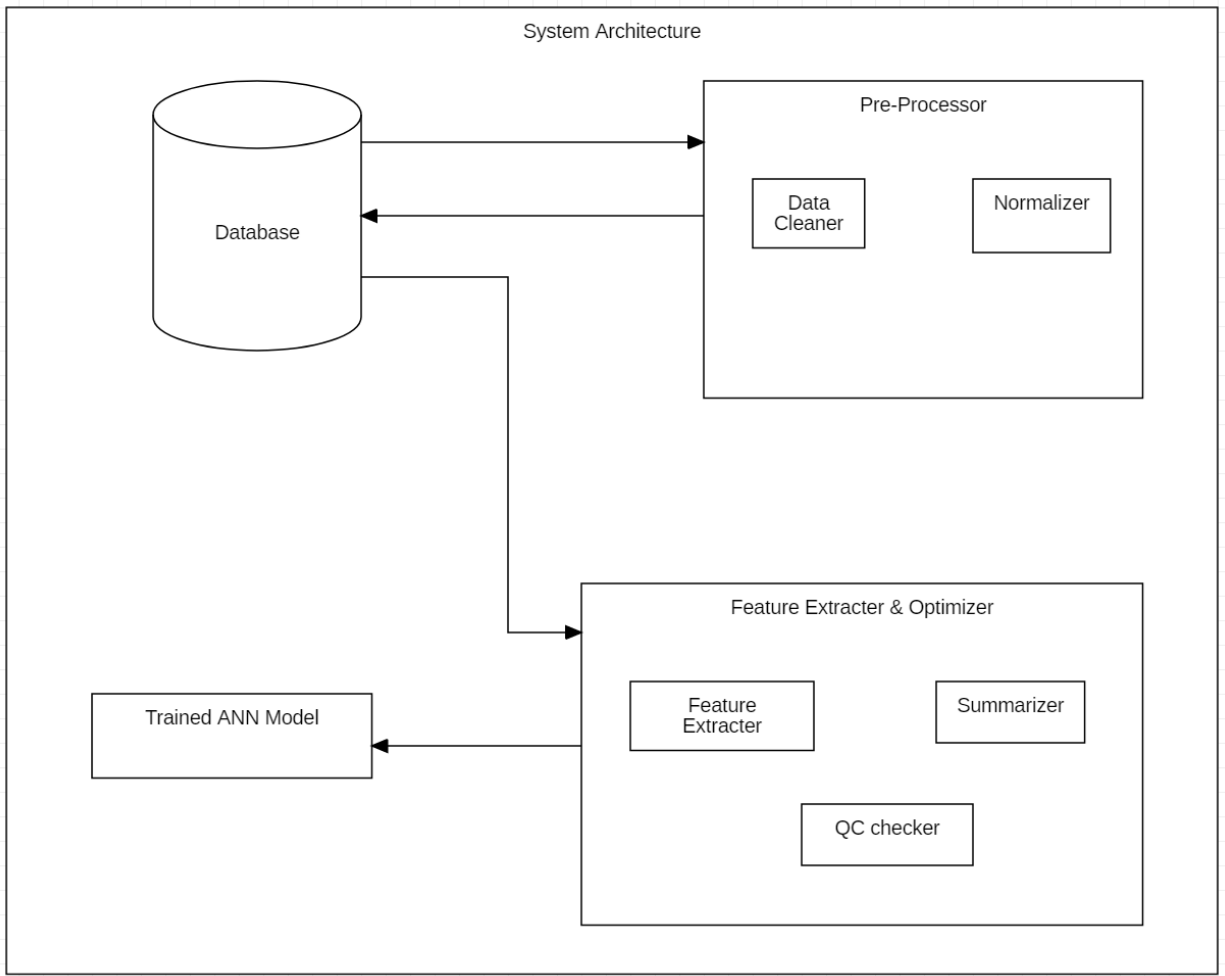


Figure 4.1: System Architecture

4.4 Use Case Diagram

The use case diagram is used to depict the interaction between actors and the system. Figure 4.2 illustrates this interaction as well the proposed functionality that the system should have.

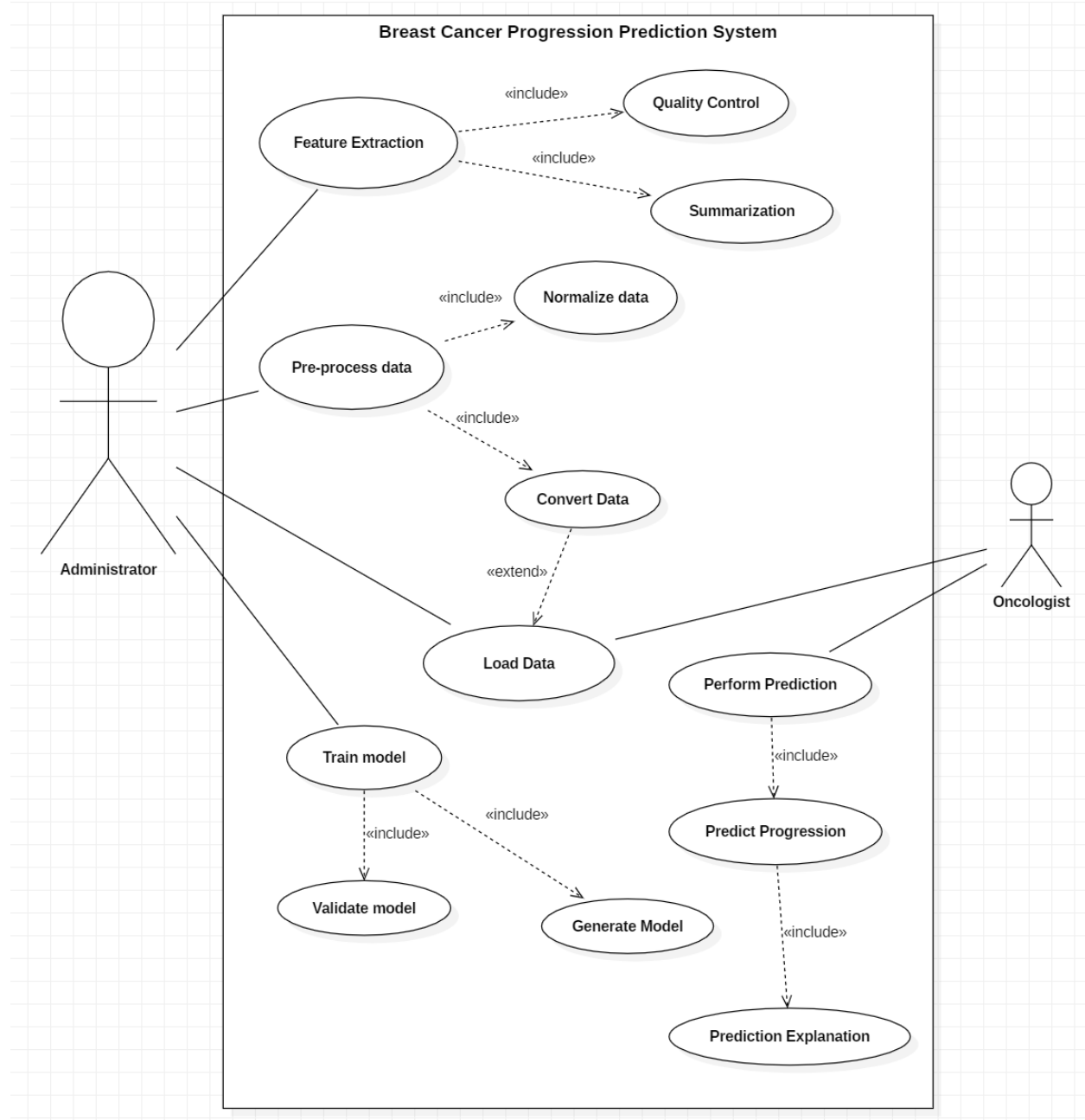


Figure 4.2: Breast Cancer Progression Prediction Use Case

4.5 System Sequence Diagram

The system sequence diagram illustrates how the primary user and the system interact. Figure 4.3 illustrates this interaction. The oncologist will first initiate the process of prediction. He/She will load the data that they intend to get prediction results for and then end the prediction. The system will then output the predicted values to show how the breast cancer will have progressed. Lastly,

the oncologist will ask for an explanation of how the prediction was arrived at and the system will output the explanation.

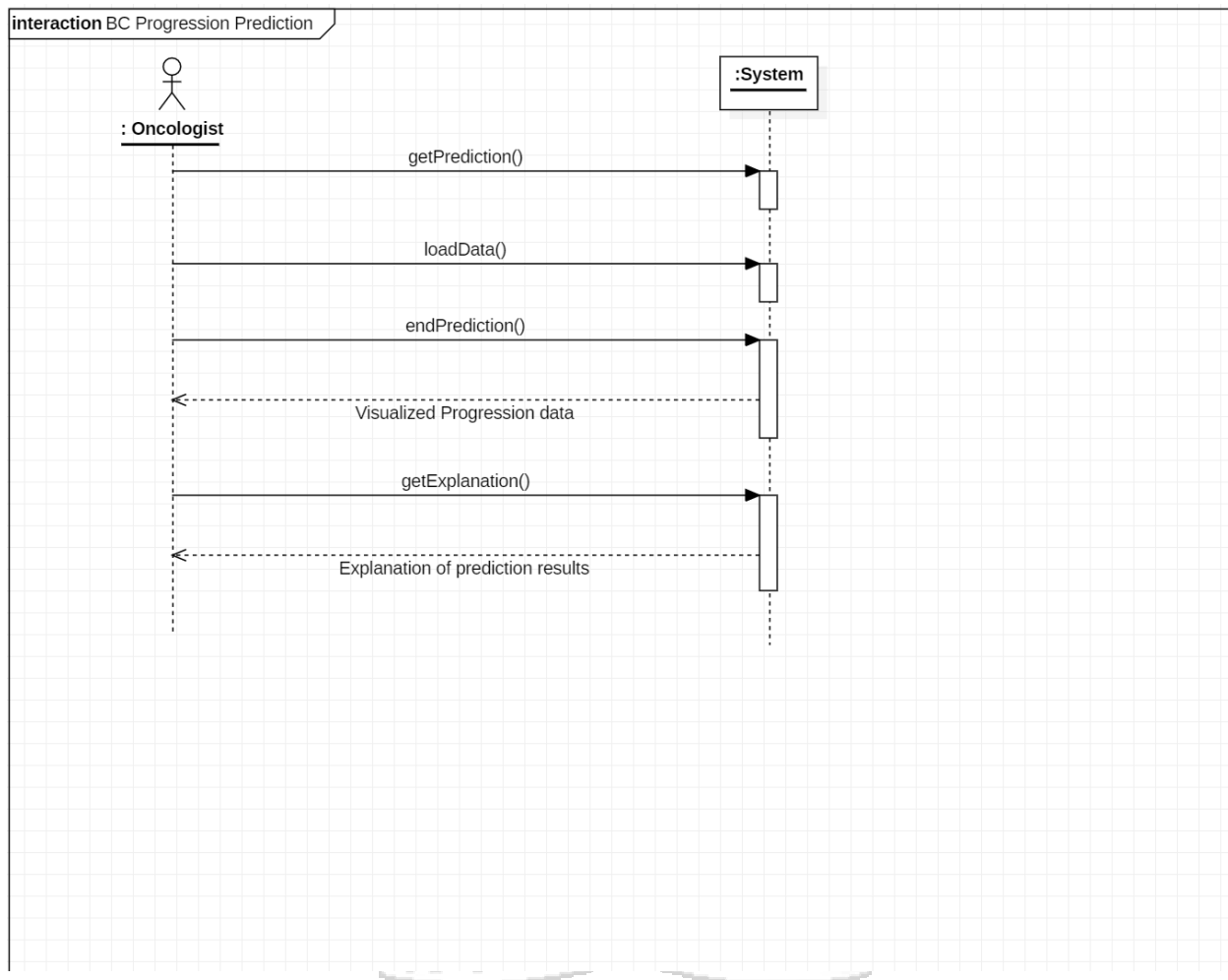


Figure 4.3: System Sequence Diagram

4.6 Sequence Diagram

The sequence diagram shown in Figure 4.4 shows the sequence of interactions between the user and the internal components of the system. The data uploaded is first processed and the processed data undergoes feature extraction. The data is then split into training and test cases and fed into the artificial neural network to create the model. The system then performs the prediction and sends it to the oncologist with the percentage accuracy. The oncologist then asks for an explanation to the prediction and the system sends the explanation as to how the prediction was obtained.

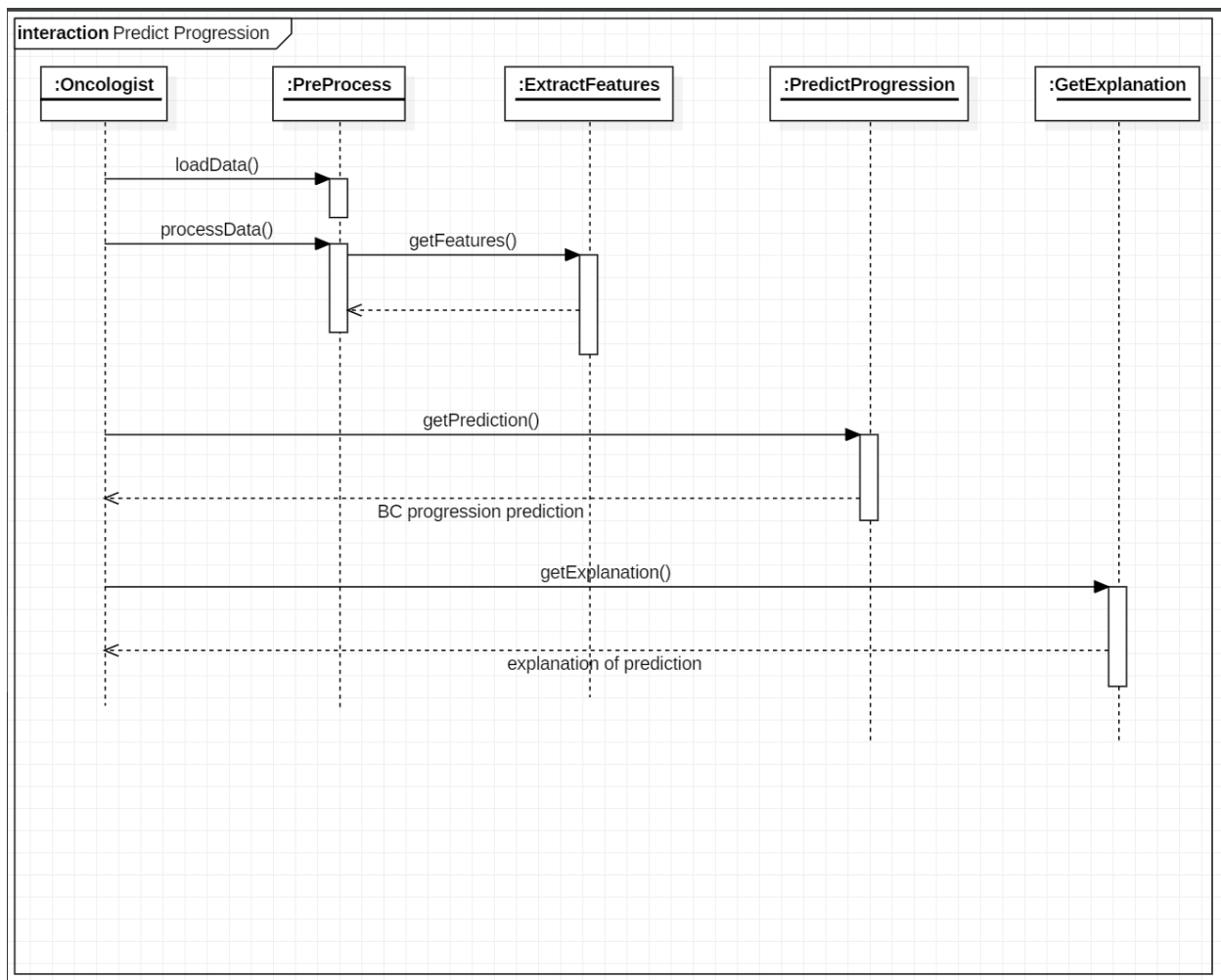


Figure 4.4: Sequence Diagram

4.7 Flow Chart Diagram

The process flow of how the training model is obtained is shown in Figure 4.5. Patient data is entered and the data is then pre-processed. The data is then tested based on the existing generated model based on the predefined number of epochs (stopping condition). If the model is not optimal, then it checks if the number of epochs that have to be run while training the model has been met. If the number of epochs has been met, the system stops.

If the number has not been met, the weights are readjusted and a new model is trained. The system repeats this process until the stopping condition has been met. The optimal model is then saved and will be used at runtime.

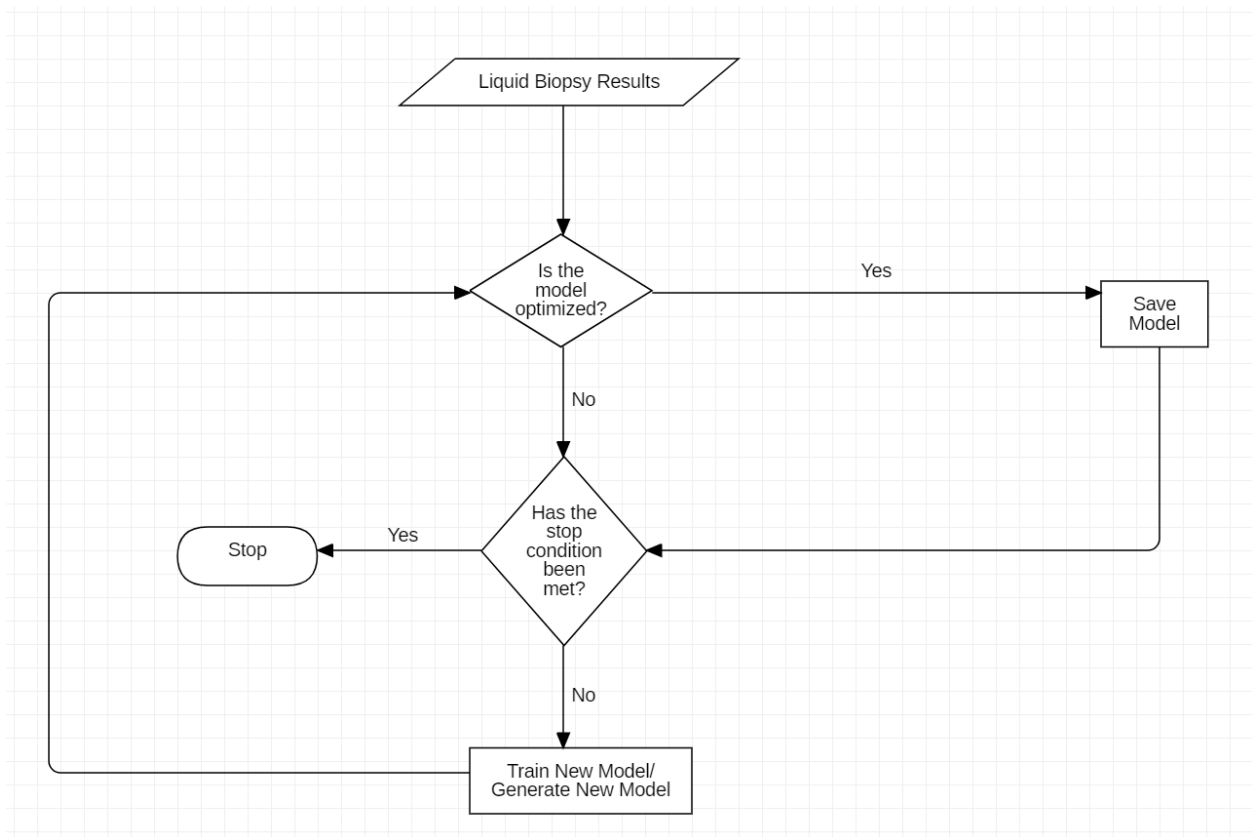
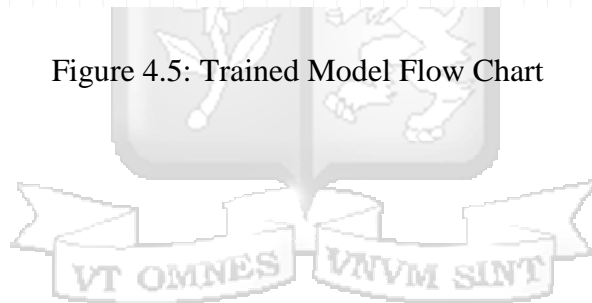


Figure 4.5: Trained Model Flow Chart



Chapter 5: Implementation

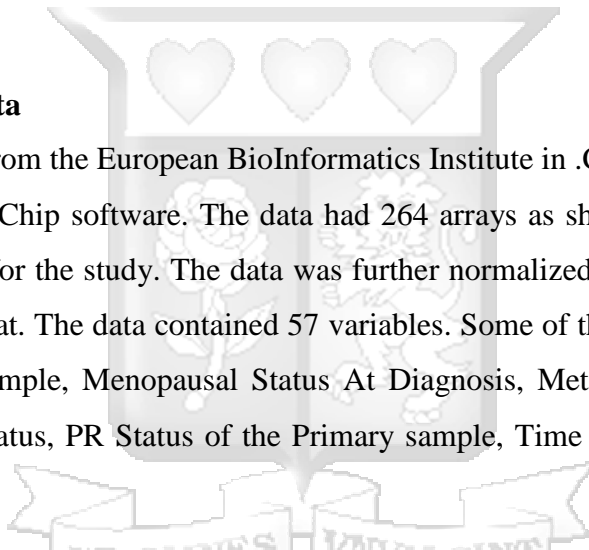
5.1 Introduction

This chapter describes the process of how the prototype was implemented, trained, tested, validated and optimized. The process starts by extracting data and normalizing that data. Pre-processing is then examined and afterwards the model is trained. The model is then validated by using the validation data set and it is optimized by adjusting its variables. Once the model is optimized, it is tested against the test data and the results evaluated.

To validate the researcher's approach, other models were implemented including different ways of feature extraction to determine the best configuration combinations. The model with the highest accuracy was chosen.

5.2 Extraction of Data

Raw data was collected from the European BioInformatics Institute in .CEL format. The raw data was extracted by using dChip software. The data had 264 arrays as shown in Figure 5.1 which were treated as samples for the study. The data was further normalized and summarized and the results saved in .xls format. The data contained 57 variables. Some of the variables included: ER Status of the Primary sample, Menopausal Status At Diagnosis, Metastatic Recurrence Time, Overall Patient HER2 Status, PR Status of the Primary sample, Time To Death as well as the Tumor stage.



```
Found 260: C:\Users\albert\Documents\Masters\Thesis\Data\CancerData\N7P1 (S0555FL0011a).CEL
Accessing 'N7P1 (S0555FL0011a).dcp' (file format 4)
Found 261: C:\Users\albert\Documents\Masters\Thesis\Data\CancerData\N7T1 (S0555FL0012).CEL
Accessing 'N7T1 (S0555FL0012).dcp' (file format 4)
Found 262: C:\Users\albert\Documents\Masters\Thesis\Data\CancerData\N8L1 (S0555FL0119).CEL
Accessing 'N8L1 (S0555FL0119).dcp' (file format 4)
Found 263: C:\Users\albert\Documents\Masters\Thesis\Data\CancerData\N8P1 (S0555FL0116).CEL
Accessing 'N8P1 (S0555FL0116).dcp' (file format 4)
Found 264: C:\Users\albert\Documents\Masters\Thesis\Data\CancerData\N8T1 (S0555FL0117).CEL
Accessing 'N8T1 (S0555FL0117).dcp' (file format 4)
```

Writing file 'C:\Users\albert\Desktop\dChip_array_summary.xls' failed. Possibly the file is open in Excel or wrong directory name

Treat all the 264 arrays as 264 samples and 264 sample groups

Figure 5.1: Extracted samples

5.3 Pre-processing

The data was pre-processed by performing normalization. The string values were converted into nominal values to be able to train the machine learning model. After normalization, the data was divided into three: training set, validation set and test set. The split was as follows: 60% was used as the training data set, 20% was used as the validation data set and the remaining 20% as the test data set. The data from the control group was excluded from pre-processing and was not considered in the eventual generation of the model. The data sets were then saved in .CSV files that would be used later in training, validating and testing the model.

5.4 Training the Model

Once the data was pre-processed, the model training could proceed. The .CSV training data file was read into the Weka platform. Before the training data was passed into the deep learning classifier, feature engineering was performed on it. A genetic search was performed on the data so as to select the best subset of attributes that would be used in the classifier to obtain a greater accuracy of prediction. The genetic algorithm parameters were configured as shown in Figure 5.2. The population size which represented the number of individuals in the training data set was set to 40. The maxGenerations which was the stopping function was set to 20 signifying that the processing would immediately stop after 20 generations have been created.



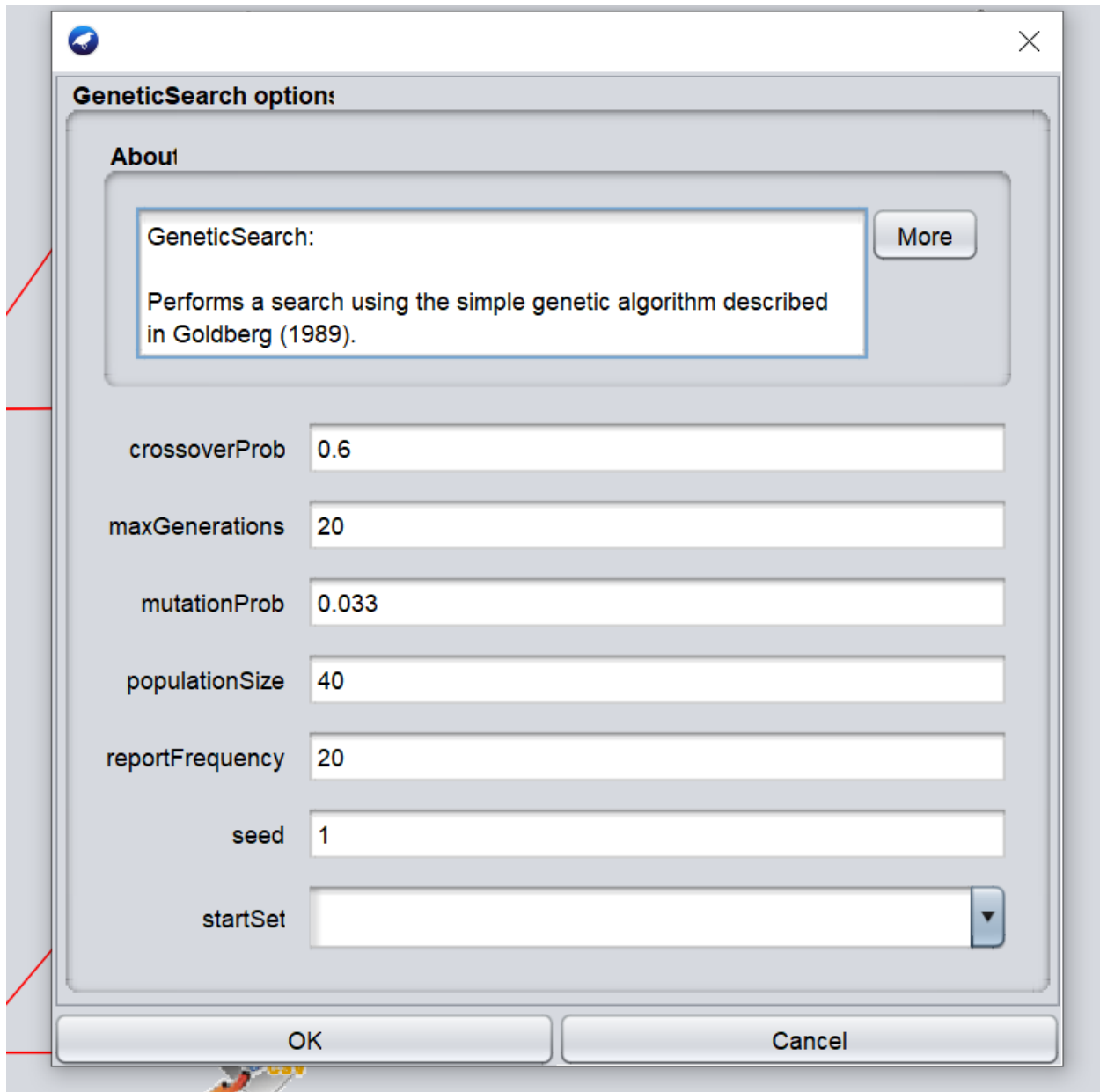


Figure 5.2: Genetic Algorithm Parameters

Once the feature engineering was completed, the data was sent to the deep learning classifier that was provided as part of the Weka DeepLearning package. The classifier was configured with the following parameters:

1. The number of epochs was set to 10.
2. The hidden layers were configured using the LeakyReLU activation function

3. The output layer was configured by using the Softmax function and the loss function used was Mean Squared Error loss function

The model was then saved so that it can be validated and optimized later. See Figure 5.3 for a diagrammatic representation of the process employed in training and saving the model.

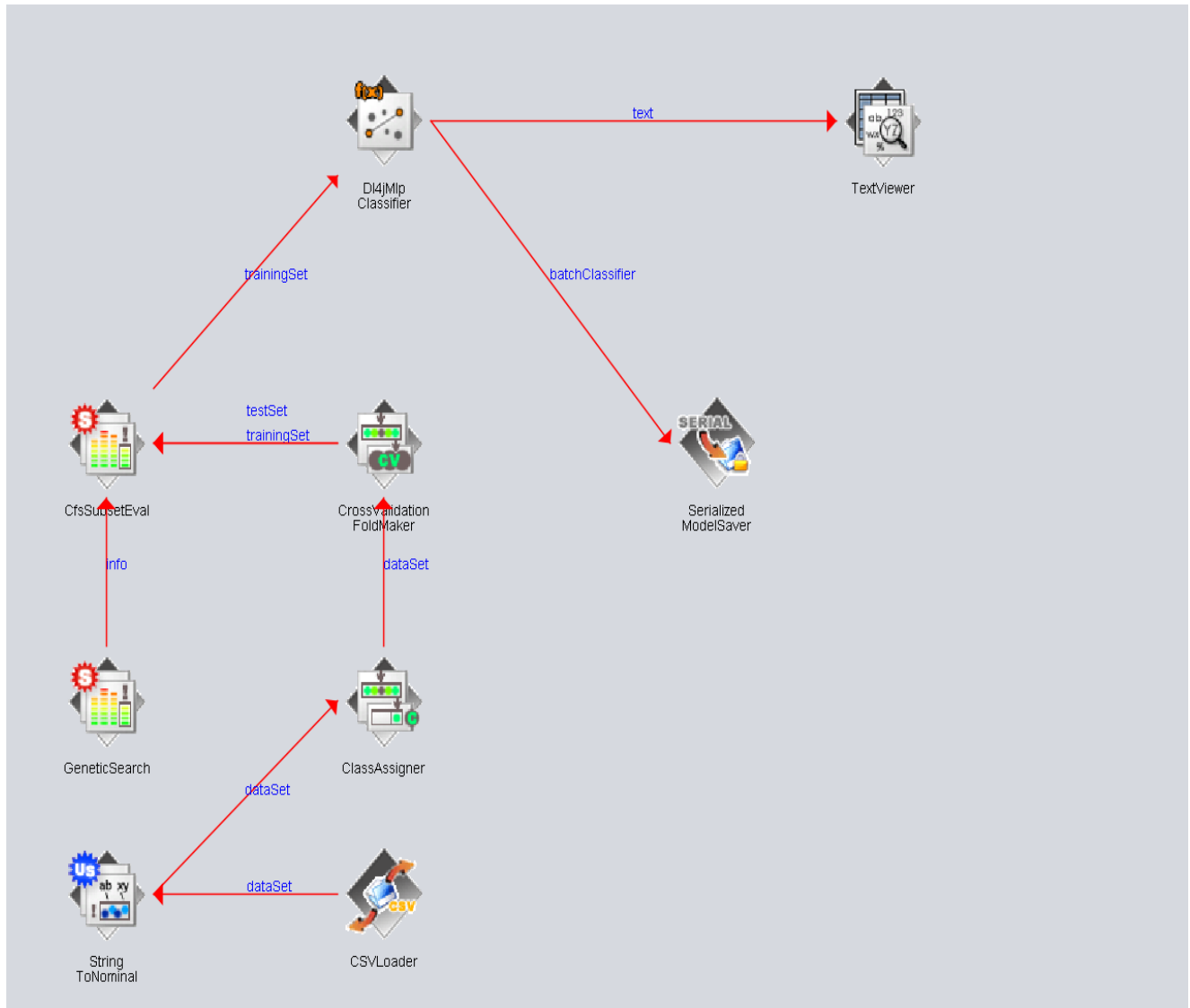


Figure 5.3: Training and saving the model

5.5 Validating the Model

After the model was trained, the validation data set was used to validate the model. 10-fold cross validation was used and the results evaluated. See Figure 5.3 for the results obtained from the cross-validation. The results from the cross-validation were evaluated to determine how fit the

model was to make predictions. The Relative Absolute Error and Root Relative Absolute Error figures depicted that the model was not accurate and had to be optimized for greater accuracy.

```
Time taken to build model: 31.82 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      331          57.4653 %
Incorrectly Classified Instances    245          42.5347 %
Kappa statistic                    0.4407
Mean absolute error                 0.0716
Root mean squared error             0.1881
Relative absolute error             71.6612 %
Root relative squared error         84.3743 %
Total Number of Instances          576
```

Figure 5.4: Cross-Validation results

5.6 Optimizing the Model

The model was optimized by using the Stochastic Gradient Descent optimization algorithm. The algorithm was chosen as it updates the parameters for each training example and is much faster and less resource intensive. Figure 5.5 shows the configurations undertaken in optimization of the model.

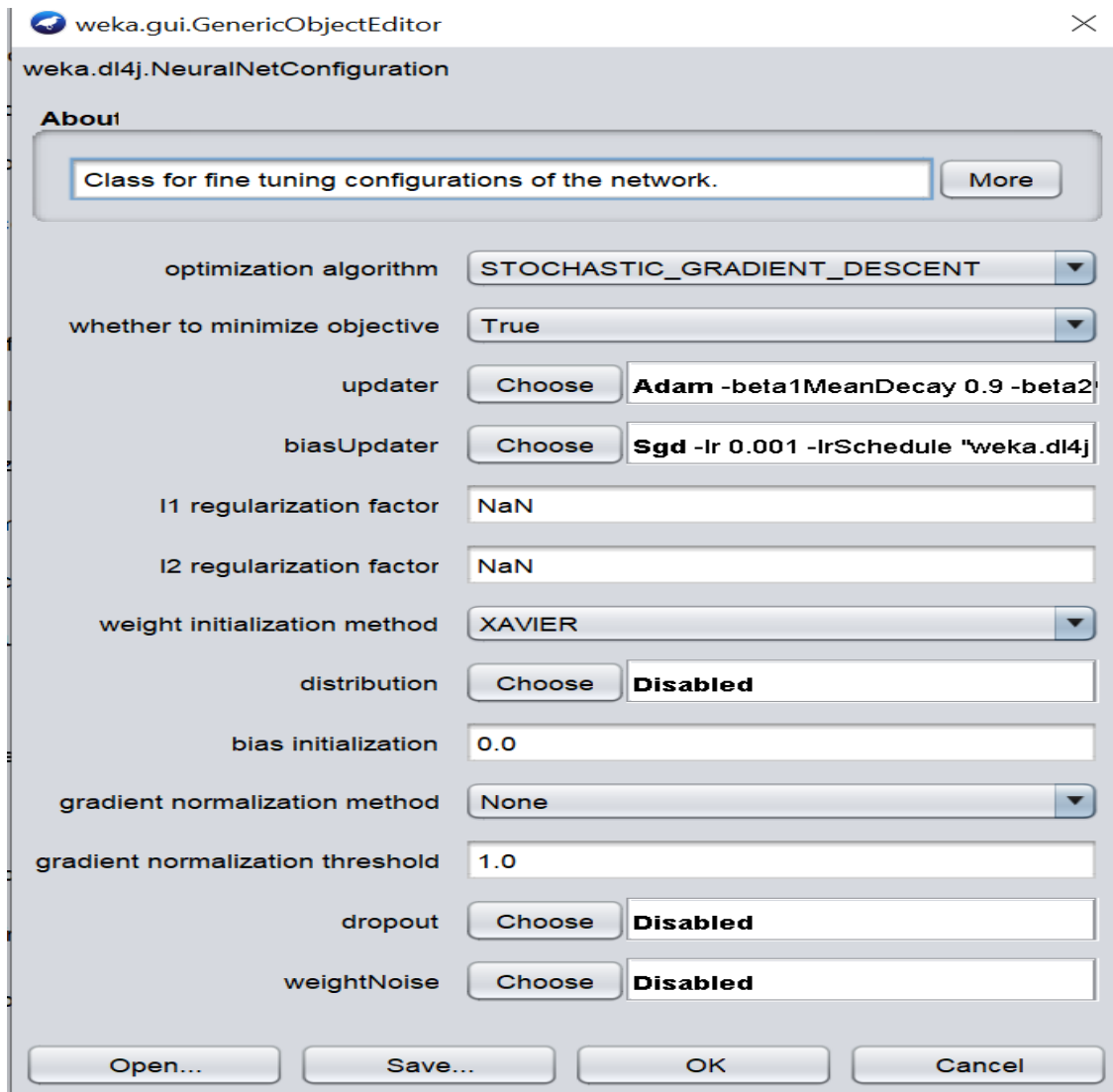


Figure 5.5: Neural network optimization

5.7 Testing the Model

After the model was optimized, the it was tested against the test data set. Figure 5.4 illustrates the results of testing the model after optimization.

Time taken to build model: 32.11 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.16 seconds

=== Summary ===

Correctly Classified Instances	94	94	%
Incorrectly Classified Instances	6	6	%
Kappa statistic	0.9203		
Mean absolute error	0.0365		
Root mean squared error	0.1037		
Relative absolute error	37.1308	%	
Root relative squared error	47.3002	%	
Total Number of Instances	100		

Figure 5.6: Test results after Optimization

5.8 Selecting the Model

The model was selected based on an evaluation from comparing results from different models built from different classifiers. The classifiers used included Naïve Bayes and the Jrip (Weka Classifier). Table 5.1 represents the different results obtained that led to selecting the best model.

Table 5.2: Model results

	Kappa	MAE	RMSE	RAE	RRSE
Naïve Bayes	0.7042	0.0319	0.1548	32.0048%	69.4307
Jrip	0.4616	0.0681	0.1846	72.713%	85.34%
Deep Learning	0.9203	0.0365	0.1037	37.1308%	47.3002%

The deep learning model was thus chosen as it gave the best degree of accuracy and model fit as compared to the other models.



Chapter 6: Discussion

6.1 Introduction

This chapter highlights the results of the research in light of the objectives set out. The main objective of the research was to create a learning prediction prototype that can predict the progression of breast cancer. A deep learning model was created after the dimensionality of the data was reduced by use of genetic algorithm. Different models were evaluated and the best model with the best performance determined as the deep learning model with genetic algorithms used to select the best subset of attributes.

6.2 Results

The aim of this experiment was to compare the performance of the deep learning model implemented with Principal Components Analysis(PCA) and without any feature engineering. The model was then evaluated and optimized and the results of the experiment are summarized in Table 6.1

Table 6.1 : Performance Comparison

Classifier	Kappa	MAE	RMSE	RAE	RRSE
Deep learning without FE	0.234	0.1806	0.4244	53.4959 %	103.3475 %
Deep learning with PCA	0.5384	0.2394	0.3276	70.9198 %	79.7767 %
Deep learning with Genetic Algorithm	0.9203	0.0365	0.1037	37.1308%	47.3002%

The results in Table 6.1 show that the deep learning model with genetic algorithm used for feature selection performs better than with model with principal components analysis and without across all the metrics.

6.3 Research Findings

The research results found out that the deep learning model enhanced with feature selection by use of genetic algorithms as the best approach to predicting breast cancer progression. The model was able to predict the Tumor stage as well as the time the cancer progresses to become metastatic with a 94% degree of accuracy. This degree of accuracy was achieved as a result of combination of factors. Firstly, the optimization of the genetic algorithm used as a feature engineering technique. Lastly, the combination of the different functions used in the input and output layers of the deep learning algorithm coupled with the loss function optimized the model and thus gave a higher accuracy of 94% as compared to the other algorithms experimented. The higher Kappa value of 0.9203 also signifies that the model's expected output is closer to the observed output signifying a better model fit.



Chapter 7: Conclusions, Recommendations and Future Work

7.1 Conclusion

The main goal of the research was to develop a learning prototype that was able to predict breast cancer progression. Breast cancer progression is identified by either the cancer progressing and affecting other parts of the body or moving from one stage to the next.

Due to the prognostic value of Circulating Cell-free DNA, biomarkers and status of the major cancer genes in it were evaluated and used to develop a deep learning model that was able to predict the tumor stage, time taken to progress as well as the time taken for metastasis to occur with an accuracy of 94%

Different machine learning algorithms have different theoretical implications based on the data under review. As a result, the various experiments with different classifiers and feature selection algorithms helped select the best model.

7.2 Recommendations

The accuracy of deep learning algorithms is highly dependent on the volume of data being studied. This poses a significant challenge especially when performing studies that involve clinical data. Secondly, the sensitivity of the data to the collection techniques employed is a major factor to consider when performing such studies. Lastly, the researcher proposes further study in use of home measures that interface with the doctors and/or care givers with predictive capabilities enabled so that medical intervention can be determined in advance of an occurrence requiring one (Tonarelli, 2017).

The researcher, therefore, recommends that such studies be done in close partnership with health institutions and/or professionals. This would ensure greater availability of data as well as a better understanding of the impact of the tools used in collecting clinical data and how to incorporate them in the model. Furthermore, there needs to be greater exploration of the ethical implications of using deep learning in a clinical setting and what measures need to be put in place to ensure standardization of applications developed for use in this setting.

7.3 Future Work

There is still a tremendous amount of work that needs to be done to understand breast cancer, how it progresses and the heterogeneity of tumors associated with it. There is greater need of making more data available to researchers in ensuring that breast cancer is understood and eventually a cure is developed.

The data that is widely available is data that has been collected by foreign bodies for patients who are predominantly Caucasian. The researcher proposes further studies to be done on breast cancer by employing technologies such as Federated learning where patients can freely share their medical data through their mobile phones and models trained by leveraging on the end users devices.



References

- Abreu, P. H., Santos, M. S., Abreu, M. H., Andrade, B. A., & Silva, D. C. (2016). Predicting Breast Cancer Recurrence Using Machine Learning Techniques: A Systematic Review. *ACM Computing Surveys*, 49(3), 1-40.
- Ahmad, F. K., & Yusoff, N. (2013). Classifying breast cancer types based on fine needle aspiration biopsy data using random forest classifier. *13th International Conference on Intelligent Systems Design and Applications* (pp. 121-125). Bangi: IEEE.
- Amidi, A., & Amidi, S. (n.d.). *Recurrent Neural Networks*. Retrieved July 28, 2019, from Stanford University: <https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-recurrent-neural-networks>
- Amrane, M., Oukid, S., Gagaoua, I., & Ensari, T. (2018). Breast cancer classification using machine learning. *Electric Electronics, Computer Science, Biomedical Engineerings' Meeting* (pp. 1-4). Istanbul: IEEE. doi:10.1109/EBBT.2018.8391453
- Bair, E., Hastie, T., Debashis, P., & Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association*, 119-137.
- Bernico, M. (2018). *Deep Learning Quick Reference : Useful Hacks for Training and Optimizing Deep Neural Networks with TensorFlow and Keras*. Packt Publishing Limited.
- Breast Cancer Statistics*. (2018). Retrieved from World Cancer Research Fund: <https://www.wcrf.org/dietandcancer/cancer-trends/breast-cancer-statistics>
- Brownlee, J. (2019, May 22). *Difference between Classification and Regression in Machine Learning*. Retrieved from Machine Learning Mastery: machinelearningmastery.com/classification-versus-regression-in-machine-learning/
- Cancer*. (2018, September 12). Retrieved from World Health Organization: <https://www.who.int/news-room/fact-sheets/detail/cancer>

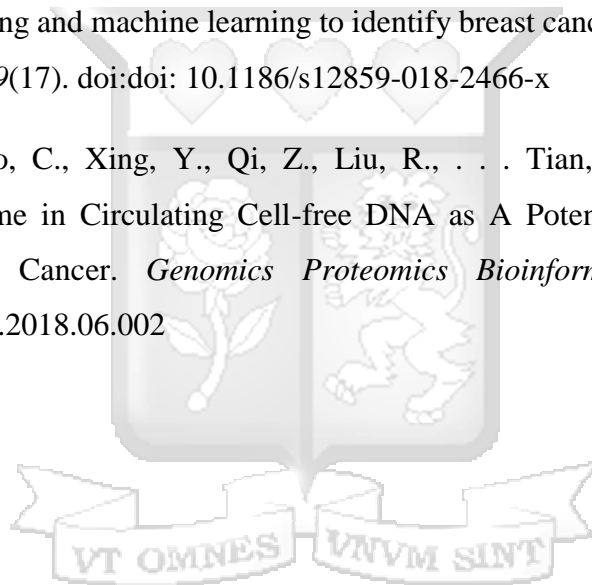
- Chaudhry, S. S., Varano, M. W., & Xu, L. (2000). Systems Research, Genetic Algorithms and Information Systems. *Systems Research and Behavioral Science*, 17, 149-162.
- Cheng, J., Cuk, K., Heil, J., Golatta, M., Schott, S., Sohn, C., . . . Surowy, H. (2017, April 24). Cell-free circulating DNA integrity is an independent predictor of impending breast cancer recurrence. *Oncotarget*, 8(33), 54537-54547.
- Computed Tomography (CT)*. (2020). Retrieved June 10, 2020, from National Institute of Biomedical Imaging and Bioengineering: <https://www.nibib.nih.gov/science-education/science-topics/computed-tomography-ct>
- Crowley, E., Di Nicolantonio, F., Loupakis, F., & Bardelli, A. (2013, July 9). Liquid biopsy: monitoring cancer-genetics in the blood. *Nature Reviews Clinical Oncology*, 10, 472–484.
- Di Meo, A., Bartlett, J., Cheng, Y., Pasic, M. D., & Yousef, G. M. (2017, April 14). Liquid biopsy: a step forward towards precision medicine in urologic malignancies. *Molecular Cancer*, 16(80).
- Doufekas, K., Achampong, Y., & Olaitan, A. (2019). Prevention of Cervical Cancer. *Uterine Cervical Cancer*, 17-29. doi:https://doi.org/ezproxy.library.strathmore.edu/10.1007/978-3-030-02701-8_2
- Hamim, M., El Moudden, I., Moutachaouik, H., & Hain, M. (2020). Decision Tree Model Based Gene Selection and Classification for Breast Cancer Risk Prediction. *Smart Applications and Data Analysis*, 165-177.
- Highsmith, J., & Cockburn, A. (2001). Agile Software Development: The Business of Innovation. *Computer*, 34(9), 120-127.
- Howley, T., Madden, M. G., O'Connell, M.-L., & Ryder, A. G. (2006). The effect of principal component analysis on machine learning accuracy with high-dimensional spectral data. *Knowledge Based Systems*, 363-370.
- HPV Information Centre. (2018, December 10). *Kenya: Human Papilloma Virus and Related Cancers, Fact sheet 2018*. Retrieved from HPV Information Centre: https://hpcvcentre.net/statistics/reports/KEN_FS.pdf

- (2015). *Kenya National Strategy for the Prevention and Control Non-Communicable Diseases 2015-2020*. Division of Non-communicable diseases. Nairobi: Ministry of Health.
- Kim, G. H., & Kim, S. H. (2019). Variable Selection for Artificial Neural Networks with Applications for Stock Price Prediction. *Applied Artificial Intelligence*, 33(1), 54-67. doi:10.1080/08839514.2018.1525850
- Lamy, J.-B., Sekar, B., Guezennec, G., Bouaud, J., & Séroussi, B. (2019). Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. *Artificial Intelligence in Medicine*, 94, 42-53. doi:https://doi.org/10.1016/j.artmed.2019.01.001
- Magnetic Resonance Imaging (MRI)*. (2020). Retrieved from National Institute of Biomedical Imaging and Bioengineering: <https://www.nibib.nih.gov/science-education/science-topics/magnetic-resonance-imaging-mri>
- Mammography*. (n.d.). Retrieved from National Institute of Biomedical Imaging and Bioengineering: <https://www.nibib.nih.gov/science-education/science-topics/mammography>
- Mani, S., Chen, Y., Li, X., Arlinghaus, L., Chakravarthy, B. A., Abramson, V., . . . Yankeelov, T. E. (2013). Machine learning for predicting the response of breast cancer to neoadjuvant chemotherapy. *Journal of the American Medical Informatics Association*, 20(4), 688-695.
- Mitchell, T. M. (1997). In T. M. Mitchell, *Machine Learning* (pp. 81-127). New York: McGraw-Hill.
- Niazi, A., & Leardi, R. (2012, April 15). Genetic algorithms in chemometrics. *Journal of Chemometrics*.
- Nuclear medicine*. (2020). Retrieved from National Institute of Biomedical Imaging and Bioengineering: <https://www.nibib.nih.gov/science-education/science-topics/nuclear-medicine>
- Oduor, J. T. (2017). *A Model for Home-Based Remote Monitoring of Asthmatic Patients*. Nairobi: Strathmore University.

- Oj, S., Lehner, J., Braun, M., & Holdenrieder, S. (2014). Circulating Cell Free DNA as Blood Based Biomarker in Breast Cancer. *Molecular Biology*, 3(120).
- Okonkwo, W. C., & Huisman, M. (2018). The Use of System Development Methodologies in the Development of Mobile Applications: Are they Worthy of Use? *2018 IEEE 42nd Annual Computer Software and Applications Conference* (pp. 278-283). Tokyo: IEEE.
- Panch, T., Szolovits, P., & Atun, R. (2018, October). Artificial intelligence, machine learning and health systems. *Journal of Global Health*.
- Pérez-Barrios, C., Nieto-Alcolado, I., Torrente, M., Jiménez-Sánchez, C., Calvo, V., Gutierrez-Sanz, L., . . . Romero, A. (2016, December). Comparison of methods for circulating cell-free DNA isolation using blood from cancer patients: impact on biomarker testing. *Translational Lung Cancer Research*, 5(6), 665–672. doi:10.21037/tlcr.2016.12.03
- Piau, A., Crissey, R., Brechemier, D., Balardy, L., & Nourhashemi, F. (2019, August). A smartphone Chatbot application to optimize monitoring of older patients with cancer. *International Journal of Medical Informatics*, 128, 18-23. doi:https://doi.org/10.1016/j.ijmedinf.2019.05.013
- Rençberoğlu, E. (2019, April 1). *Fundamental Techniques of Feature Engineering for Machine Learning*. Retrieved from Towards Data Science: <https://towardsdatascience.com/feature-engineering-for-machine-learning-3a5e293a5114>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why Should I Trust You?” Explaining the Predictions of any Classifier. *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 1135-1144). San Francisco, California.
- Rossi, G., Mu, Z., Rademaker, A. W., Austin, L. K., Strickland, K. S., Costa, R. L., . . . Cristofanilli, M. (2018). Cell-Free DNA and Circulating Tumor Cells: Comprehensive Liquid Biopsy Analysis in Advanced Breast Cancer. *Clinical Cancer Research*, 24(3), 560-568.

- Shaw, J. A., Page, K., Blighe, K., Hava, N., Guttery, D., Ward, B., . . . Ruangpratheep, C. (2012). Genomic analysis of circulating cell-free DNA infers breast cancer dormancy. *Genome Research*, 220-231.
- Shulman, L., Willett, W., Sievers, A., & Knaul, F. (2010). Breast Cancer in Developing Countries: Opportunities for Improved Survival. *Journal of Oncology*.
- Sobhani, N., Generali, D., Zanconati, F., Bortul, M., & Scaggiante, B. (2018, April). Cell-free DNA integrity for the monitoring of breast cancer: Future perspectives? *World journal of clinical oncology*, 9(2), 26-32. doi:10.5306/wjco.v9.i2.26
- Solomatine, D., See, L. M., & Abrahart, R. J. (2009). Data-Driven Modelling: Concepts, Approaches and Experiences. In D. Solomatine, L. M. See, & R. J. Abrahart, *Practical Hydroinformatics* (pp. 17-30). Berlin, Heidelberg: Springer.
- Tabl, A. A., Alkhateeb, A., ElMaraghy, W., Rueda, L., & Ngom, A. (2019). A Machine Learning Approach for Identifying Gene Biomarkers Guiding the Treatment of Breast Cancer. *Frontiers in Genetics*.
- The Aga Khan University Hospital. (2019, January 3). *Cervical cancer a leading cause of death among women*. Retrieved from The Aga Khan University Hospital, Nairobi: <https://hospitals.aku.edu/nairobi/AboutUs/News/Pages/cervical-cancer.aspx>
- Tonarelli, L. (2017, February 22). *Innovative ways digital health can help manage cancer*. Retrieved October 12, 2020, from Philips: <https://www.philips.com/a-w/about/news/archive/standard/news/press/2020/20200828-philips-expands-its-dedicated-cardiovascular-ultrasound-offering-by-launching-affiniti-cvx-for-increased-productivity.html>
- Tseng, Y., Huang, C., Wen, C., Lai, P. Y., Wu, M., Sun, Y., . . . Lu, J. J. (2019). Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data with machine learning technologies. *International Journal of Medical Informatics*, 79-86.
- Types of Cancer Treatment*. (2020). Retrieved November 9, 2020, from National Cancer Institute: <https://www.cancer.gov/about-cancer/treatment/types>

- Vanneschi, L., Farinaccio, A., Giacobini, M., Mauri, G., Antoniotti, M., & Provero, P. (2010). Identification of Individualized Feature Combinations for Survival Prediction in Breast Cancer: A Comparison of Machine Learning Techniques. *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics* (pp. 110-121). Berlin: Springer, Berlin, Heidelberg.
- Yu, D., Tong, Y., Guo, X., Feng, L., Jiang, Z., Ying, S., . . . Lou, J. (2019). Diagnostic Value of Concentration of Circulating Cell-Free DNA in Breast Cancer: A Meta-Analysis. *Frontiers in Oncology*.
- Zeng, Z., Espino, S., Roy, A., Li, X., Khan, S. A., Clare, S. E., . . . Luo, Y. (2018). Using natural language processing and machine learning to identify breast cancer local recurrence. *BMC Bioinformatics*, 19(17). doi:doi: 10.1186/s12859-018-2466-x
- Zhang, J., Han, X., Gao, C., Xing, Y., Qi, Z., Liu, R., . . . Tian, X. (2018, June 16). 5-Hydroxymethylome in Circulating Cell-free DNA as A Potential Biomarker for Non-small-cell Lung Cancer. *Genomics Proteomics Bioinformatics*, 16(3), 187-199. doi:10.1016/j.gpb.2018.06.002



Appendix: Originality Report



Predicting Breast Cancer Progression by using Cell-free DNA.docx

ORIGINALITY REPORT

12%	9%	6%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.researchgate.net Internet Source	1%
2	hdl.handle.net Internet Source	1%
3	Submitted to Universiti Sains Islam Malaysia Student Paper	<1%
4	Submitted to Birkbeck College Student Paper	<1%
5	ecancer.org Internet Source	<1%
6	www.wjgnet.com Internet Source	<1%
7	Submitted to Royal Melbourne Institute of Technology Student Paper	<1%
8	Submitted to University of Central Lancashire Student Paper	<1%